

# Brisk guide to Mathematics

**Jan Slovák**

and

**Michal Bulant, Ioannis Chrysikos, Martin Panák**

with help of

**Ray Booth, Vladimír Ejov, Radek Suchánek, Vojtěch Žádník, ...**

Brno, 2024

Works on this text were financially supported by the project NPO MUNI 3.2.1 DataAnalytics NPO-MUNI-MSMT-16606/2022



Financováno  
Evropskou unií  
NextGenerationEU



NÁRODNÍ  
PLÁN OBNOVY

MSMT  
MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

**Authors:**

Michal Bulant  
Ioannis Chrysikos  
Martin Panák  
Jan Slovák

**With further help of:**

Ray Booth  
Vladimir Ejev  
Radek Suchánek  
Vojtěch Žádník

**Graphics and illustrations:**

Petra Rychlá

## Contents – practice

Chapter 1. Initial warmup	4
A. Numbers and functions	4
B. Difference equations	12
C. Combinatorics	16
D. Probability	20
E. Plane geometry	28
F. Relations and mappings	43
G. Additional exercises for the whole chapter	47
Chapter 2. Elementary linear algebra	83
A. Vectors and matrices	83
B. Determinants	100
C. Vector spaces and linear mappings	107
D. Properties of linear mappings	133
E. Additional exercises for the whole chapter	142
Chapter 3. Linear models and matrix calculus	193
A. Linear optimization	193
B. Recurrence relations	204
C. Models of growth and iterated processes	210
D. More matrix calculus	224
E. Matrix decompositions	248
F. Additional exercises for the whole chapter	262
Chapter 4. Analytic geometry	319
A. Affine geometry	319
B. Euclidean geometry	327
C. Geometry of quadratic forms	341
D. Further exercise on this chapter	361
Chapter 5. Establishing the ZOO	368
A. Polynomial interpolation	368
B. Real numbers and limit processes	380
C. Derivatives	407
D. Infinite sums and power series	428
E. Additional exercises for the whole chapter	444
Chapter 6. Differential and integral calculus	508
A. Derivatives of higher orders	508
B. Integration	534
C. Sequences, series and limit processes	566
D. Additional exercises for the whole chapter	577
Chapter 7. Continuous tools for modelling	631
A. Fourier series	631
B. Integral operators and Fourier Transform	654
C. Metric spaces	667
D. Additional exercises to the whole chapter	686

## Contents – theory

Chapter 1. Initial warmup	4
1. Numbers and functions	4
2. Difference equations	10
3. Combinatorics	14
4. Probability	18
5. Plane geometry	26
6. Relations and mappings	41
Chapter 2. Elementary linear algebra	83
1. Vectors and matrices	83
2. Determinants	96
3. Vector spaces and linear mappings	107
4. Properties of linear mappings	128
Chapter 3. Linear models and matrix calculus	193
1. Linear optimization	193
2. Difference equations	202
3. Iterated linear processes	211
4. More matrix calculus	221
5. Decompositions of the matrices and pseudoinversions	244
Chapter 4. Analytic geometry	319
1. Affine and Euclidean geometry	319
2. Transformations	338
3. Geometry of quadratic forms and quadrics	342
4. Projective geometry	349
Chapter 5. Establishing the ZOO	368
1. Polynomial interpolation	368
2. Real numbers and limit processes	379
3. Derivatives	403
4. Infinite sums and power series	417
Chapter 6. Differential and integral calculus	508
1. Differentiation	508
2. Integration	529
3. Sequences, series and limit processes	557
Chapter 7. Continuous tools for modelling	631
1. Fourier series	631
2. Integral operators and Fourier transform	654
3. Metric spaces	666
Chapter 8. Calculus with more variables	709
1. Functions and mappings on $\mathbb{R}^n$	709
2. Integration for the second time	747
3. Differential equations	761

Chapter 8. Calculus with more variables	709	Chapter 9. Continuous models – further selected topics	802
A. Multivariate functions	709	1. Exterior differential calculus and integration	802
B. The topology of $E_n$	711	2. Remarks on Partial Differential Equations	827
C. Limits and continuity of multivariate functions	713	3. Remarks on Variational Calculus	858
D. Tangent lines, tangent planes, graphs of multivariate functions	715	4. Complex Analytic Functions	873
E. Taylor polynomials	722	Chapter 10. Statistics and probability theory	901
F. Extrema of multivariate functions	723	1. Descriptive statistics	901
G. Implicitly given functions and mappings	728	2. Probability	913
H. Constrained optimization	730	3. Mathematical statistics	959
I. Volumes, areas, centroids of solids	742	Chapter 11. Elementary number theory	978
J. First-order differential equations	759	1. Fundamental concepts	978
K. Practical problems leading to differential equations	768	2. Primes	982
L. Higher-order differential equations	770	3. Congruences and basic theorems	987
M. Applications of the Laplace transform	777	4. Solving congruences and systems of them	1000
N. Numerical solution of differential equations	780	5. Diophantine equations	1016
O. Additional exercises to the whole chapter	795	6. Applications – calculation with large integers, cryptography	1020
Chapter 9. Continuous models – further selected topics	802	Chapter 12. Algebraic structures	1040
A. Exterior differential calculus	802	1. Posets and Boolean algebras	1040
B. Applications of Stoke's theorem	802	2. Elements of Logic	1055
C. Equation of heat conduction	808	3. Polynomial rings	1073
D. Partial Differential Equations	809	4. Groups, rings, and fields	1086
E. Variational Problems	830	5. Coding theory	1110
F. Complex analytic functions	830	6. Systems of polynomial equations	1117
G. Additional exercises to the whole chapter	897	Chapter 13. Combinatorial methods, graphs, and algorithms	1140
Chapter 10. Statistics and probability methods	901	1. Elements of Graph theory	1140
A. Dots, lines, rectangles	901	2. A few graph algorithms	1166
B. Visualization of multidimensional data	910	3. Remarks on Computational Geometry	1188
C. Classical and conditional probability	912	4. Remarks on more advanced combinatorial calculations	1206
D. What is probability?	920		
E. Random variables, density, distribution function	922		
F. Expected value, correlation	932		
G. Transformations of random variables	937		
H. Inequalities and limit theorems	939		
I. Testing samples from the normal distribution	944		
J. Linear regression	953		
K. Bayesian data analysis	955		
L. Processing of multidimensional data	960		
Chapter 11. Number theory	978		
A. Basic properties of divisibility	978		
B. Prime numbers	981		
C. Congruences	985		
D. Solving congruences	997		
E. Diophantine equations	1012		
F. Primality tests	1015		
G. Encryption	1019		
H. Additional exercises to the whole chapter	1038		
Chapter 12. Algebraic structures	1040		
A. Boolean algebras and lattices	1040		
B. Rings	1052		
C. Polynomial rings	1053		

D. Rings of multivariate polynomials	1058
E. Algebraic structures	1063
F. Groups	1065
G. Burnside's lemma	1084
H. Codes	1087
I. Extension of the stereographic projection	1094
J. Elliptic curves	1095
K. Gröbner bases	1098
Chapter 13. Combinatorial methods, graphs, and algorithms	
A. Fundamental concepts	1140
B. Fundamental algorithms	1149
C. Minimum spanning tree	1158
D. Flow networks	1160
E. Classical probability and combinatorics	1164
F. More advanced problems from combinatorics	1169
G. Probability in combinatorics	1171
H. Combinatorial games	1178
I. Generating functions	1180
J. Additional exercises to the whole chapter	1220
Index	1228

## Preface

This textbook is a followup of the Czech course material “Matematika drsně a svižně”, reflecting many years of lecturing Mathematics at the Faculty of Informatics at the Masaryk University in Brno. Their programme required introduction to genuine mathematical thinking and precision, but within a quite limited time-frame for lectures. This endeavor was undertaken by Jan Slovák and Martin Panák since 2004, with further collaborators joining later. Our goal was to cover seriously, but quickly, about as much of mathematical methods as usually seen in bigger courses in the classical Science and Technology programmes. At the same time, we did not want to give up the completeness and correctness of the mathematical exposition. We wanted to introduce and explain more demanding parts of Mathematics together with elementary explicit examples how to use the concepts and results in practice. But we did not want to decide how much of theory or practice the reader should enjoy and in which order.

Now, we have tried to accommodate all the above features in one book providing a more or less complete account on basics of Mathematics, as taught in typical Mathematics Bachelor programmes, but in a way relevant to the coming era of computational powers and artificial intelligence resources, including the GPT and other chatbots available widely. Thus, we chose the two column format of the textbook, where the theoretical explanation on one side and the practical procedures and exercises on the other side are split. Moreover, we suppose the practice learning will be heavily supported by Computer Aided Mathematic tool, we chose mainly Sage in our illustrations. This way, we want to encourage and help the readers to find their own way. Either to go through the examples and algorithms first (perhaps with the help of Sage or GPT-like chatbots), and then to come to more serious thinking on why the things work, or the other way round.

We also hope to overcome the usual stress of the readers horrified by the amount of the stuff. With our text, they are not supposed to read through the book in a linear order. On the opposite, the readers should enjoy browsing through the text and finding their own thrilling paths through the new mathematical landscapes.

In both columns, we intend to present rather standard exposition of basic Mathematics, but focusing on the essence of the concepts and their relations. The exercises are addressing simple mathematical problems but we also try to show the exploitation of mathematical models in practice as much as possible.

We are aware that the text is written in a very compact and non-homogeneous way. A lot of details are left to readers, in particular in the more difficult paragraphs, while we try to provide a lot of simple intuitive explanation when introducing new concepts or formulating important theorems. Similarly, the examples display the variety from very simple ones to those requesting independent thinking.

We would very much like to help the reader:

- to formulate precise definitions of basic concepts and to prove simple mathematical results;
- to perceive the meaning of roughly formulated properties, relations and outlooks for exploring mathematical tools;
- to understand the instructions and algorithms underlying mathematical models and to appreciate their usage.

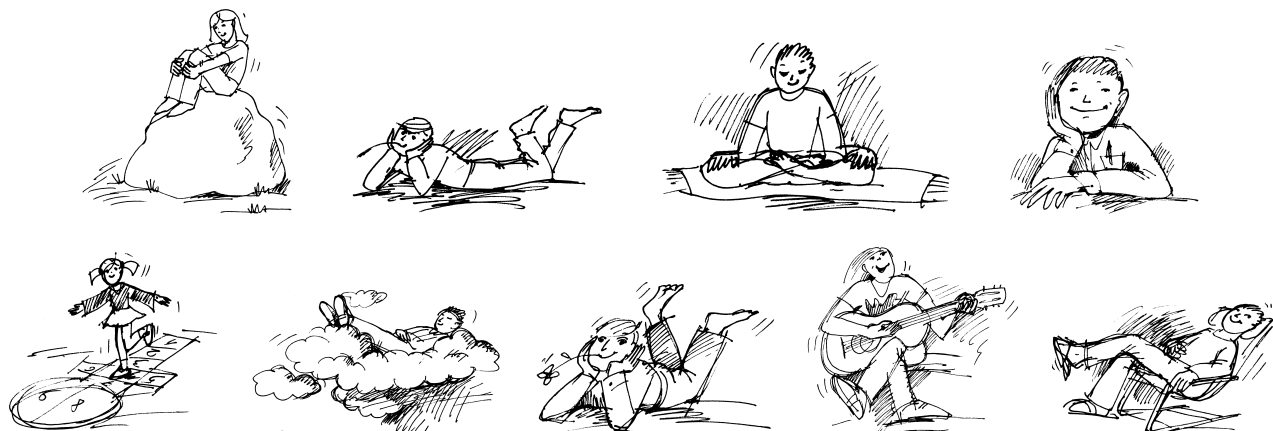
These goals are ambitious and there are no simple paths reaching them without failures on the way. This is one of the reasons why we come back to basic ideas and concepts several times with growing complexity and width of the discussions. Of course, this might also look chaotic but we very much hope that this approach gives a better chance to those who will persist in their efforts. We also hope, this textbook should be a perfect beginning and help for everybody who is ready to think and who is ready to return back to earlier parts again and again.

To make the task simpler and more enjoyable, we have added what we call "emotive icons". We hope they will spirit the dry mathematical text and indicate which parts should be read more carefully, or better left out in the first round. They should work as a sort of switches, so when loosing the ground, the reader is advised to find the next appealing icon and go on.

The usage of the icons follows the feelings of the authors and we tried to use them in a systematic way. We hope the readers will assign the meaning to icons individually. Roughly speaking, there are icons indicating complexity, difficulty etc.:



Further icons indicate unpleasant technicality and need of patience, or possible entertainment and pleasure:



Some icons are related to feelings when solving problems and appear mainly in the practical column:



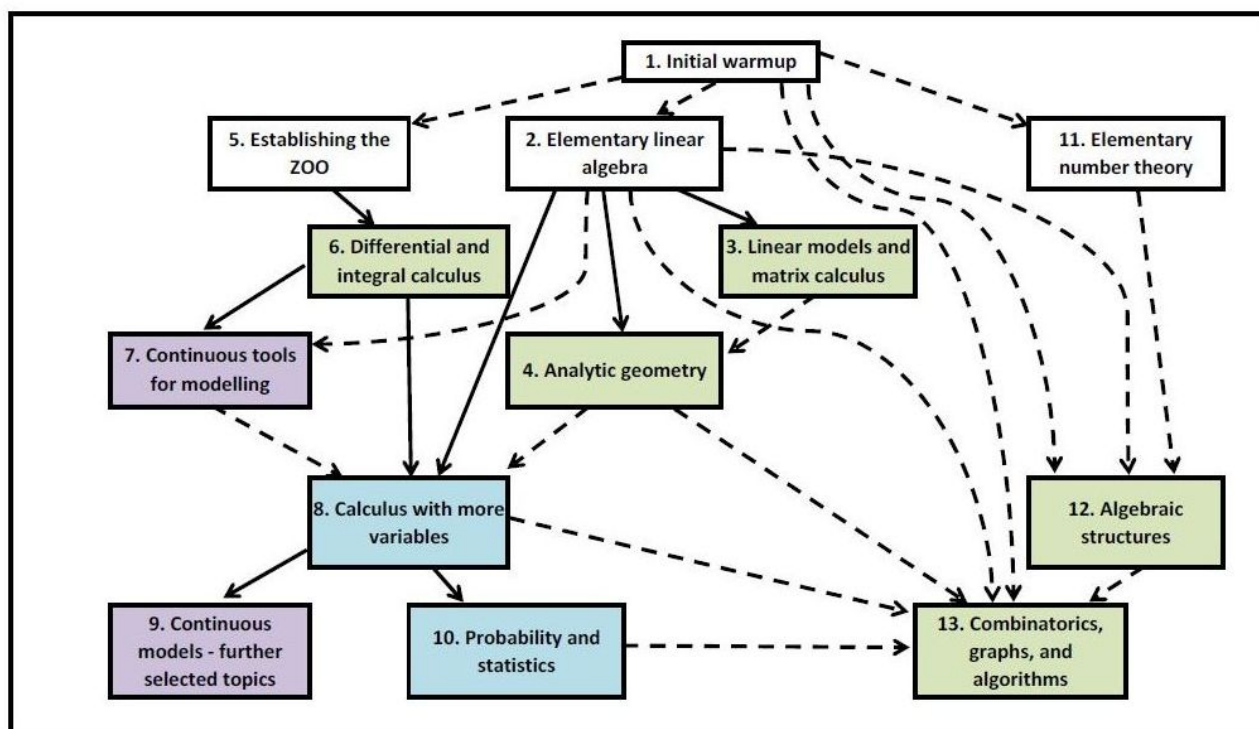
The practical column with the solved problems and exercises should be readable nearly independently of the theory. Without the ambition to know the deeper reasons why the algorithms work, it should be possible to read mainly just this column. In order to help such readers, some definitions and descriptions in the theoretical text are marked in order to catch the eyes easily when reading the exercises. The exercises and theory are partly coordinated to allow jumping there and back, but the links are not tight. The numbering in the two columns is distinguished by using the different numberings of sections, i.e., those like 1.2.1 belong to the theoretical column, while 1.A.14 points to the practical column. The equations are numbered within subsections and their quotes include the subsection numbers if necessary.

In general, our approach stresses the fact that the methods of the so called discrete Mathematics seem to be more important for mathematical models nowadays. They seem also simpler to get perceived, supported by the computational tools, and grasped.

However, the continuous methods are strictly necessary too. First of all, the classical continuous mathematical analysis is essential for understanding of convergence and robustness of computations. It is hard to imagine how to deal with error estimates and computational complexity of numerical processes without it. Moreover, the continuous models are often the efficient and effectively computable approximations to discrete problems coming from practice.

As usual with textbooks, there are numerous figures completing the exposition. We very much advise the readers to draw their own pictures whenever necessary, in particular in the later chapters, where we provide only a few.

The rough structure of the book and the dependencies between its chapters are depicted in the diagram below. The darker the color is, the more demanding is the particular chapter (or at least its essential parts). In particular, the chapters 7 and 9 include a lot of material which would perhaps not be covered in the regular course activities or required at exams in great detail. The solid arrows mean strong dependencies, while the dashed links indicate only partial dependencies. In particular, the textbook could support courses starting with any of the white boxes, i.e. aiming at standard linear algebra and geometry (chapters 2 through 4), discrete chapters of mathematics (11 through 13), and the rudiments of Calculus (5, 6, 8).



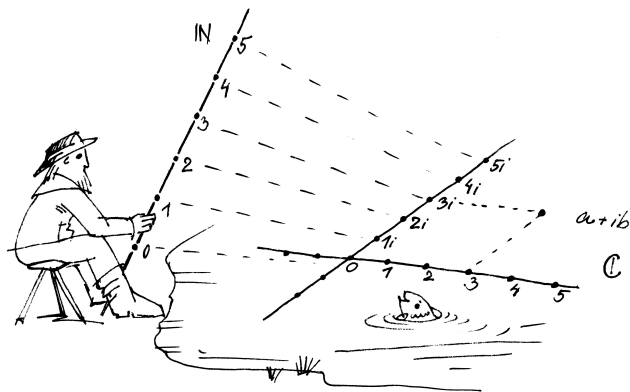
All topics covered in the book are now included (with more or less details) in teaching within our Mathematics Minor programme, complemented by numerical seminars. In this block of four courses, the first semester covers chapters 1 and 2 and selected topics from chapters 3 and 4. The second semester essentially includes chapters 5, 6, and 7. The third semester is now split into two parts. The first one is covered by chapter 8 (with only a few glimpses towards the more advanced topics from chapter 9), while the rest of the semester is devoted to the rudiments of the graph theory in chapter 13. The last semester provides large parts of the chapters 11 through 13.

This textbook project got a new impuls when developing the main background material for a new professional programme Data Analytics, a cross-disciplinary endeavor aiming at formal education of top-tier young programmers in the era of GPT and further AI resources. There, the two elementary maths courses in the first semester are covered by Chapters 1-7, while in the second semester, another two courses are based on Chapters 8-9, and 13. Finally one course in the third semester is covered by Chapters 11-12. Two more courses are devoted to Probability and Statistics and they already go much beyond the Chapter 10 here.



## Initial warmup

“value, difference, position”  
– what it is and how to comprehend it?



### A. Numbers and functions

People mostly think that “Mathematics” equals “counting”. Indeed, we start the introduction with number systems including integers, rationals, reals, and complex numbers, denoted by  $\mathbb{Z}$ ,  $\mathbb{Q}$ ,  $\mathbb{R}$  and  $\mathbb{C}$ , respectively. Counting is based on functions, and numbers are their arguments and values. We approach these elementary and well known concepts now in order to invoke the more realistic view at Mathematics as a way of thinking. The reader is always advised to keep an eye at the right-hand column if the concepts and methods get unclear. In the next pages, we also use basic properties of numbers which we carefully discuss only in Chapter 12 (e.g., the unique decomposition of integers into products of primes, etc.).

As soon as mathematical tasks approach counting, it can be easily done in computer aided mathematics software in a few lines of code. Therefore, our exposition will often include Sage cells, but we avoid presenting preliminaries on programming with Sage. Our experience indicates that the Sage (Python based) interface is user-friendly and even unfamiliar readers should feel comfortable with basic coding in Sage really fast.

**1.A.1.** Integers are *discrete*, in the sense that for any  $m \in \mathbb{Z}$ , there is no integer between  $m$  and  $m + 1$ . Show with an example that this fails for each couple of rational numbers.

The goal of this first chapter is to introduce the reader to the fascinating world of mathematical thinking.

The name of this chapter can be also understood as an encouragement for patience. Even the simplest tasks and ideas are easy only for those who have already seen similar ones.

We start with the simplest thing: numbers.

They will also serve as the first example of how mathematical objects and theories are built. The entire first chapter will become a quick tour through various mathematical landscapes (including germs of analysis, combinatorics, probability, geometry, and the language of Mathematics itself).

Perhaps sometimes our definitions and ideas will look too complicated and not practical enough. The simpler the objects and tasks are, the more difficult the mastering of depth and all nuances of the relevant tools and procedures might be. We shall come back to all of the notions again and again in the further chapters and hopefully this will be the crucial step in the ultimate understanding.

Thus the advice: do not worry if you find some particular part of the exposition too formal or otherwise difficult – come back later for another look.

### 1. Numbers and functions

Since the dawn of time, people want to know “how much” about something they have, or “how much” is something worth, “how long” will a particular task take, etc. The answer for such ideas is usually some kind of “number”. We consider something to be a number, if it behaves according to the usual rules – either according to all the rules we accept, or maybe only to some of them. For instance, the result of multiplication does not depend on the order of multiplicands. We have the number zero whose addition to another number does not change the result. We have the number one whose product with another number does not change the result. And so on.

The simplest example of numbers are the positive integers which we denote  $\mathbb{Z}^+ = \{1, 2, 3, \dots\}$ . The natural numbers consist of either just the positive integers, or the positive integers together with the number zero. The number zero is either considered to be a natural number, as is usual in computer science, or not a natural number as is usual in some other contexts.

**Solution.** Let  $p = a/b$  and  $q = c/d$  be two rationals, where  $a, b, c, d \in \mathbb{Z}$ , and suppose for instance that  $p < q$ . Consider the average  $\alpha$  of  $p, q$ , i.e.,  $\alpha = \frac{1}{2}(p + q)$ . Then  $\alpha = \frac{ad+cb}{2bd}$ , and this is a rational number as the quotient of the integers  $ad+cb$  and  $2bd$ . Now, adding  $p$  to both sides of the inequality  $p < q$  we get  $2p < p + q$ , whereas adding  $q$  to both sides gives  $p + q < 2q$ . Hence we have  $2p < p + q < 2q$  and so  $p < \alpha < q$ .  $\square$

**1.A.2.**

**1.A.3.** Show that the integer 2 does not have a rational square root (hint: think about the number of copies of 2 appearing in the decompositions of integers).  $\circ$

**1.A.4.** An irrational number is a real number which cannot be written as the ratio of two integers. Using 1.A.3, prove that between any two different rational numbers there is an irrational number.  $\circ$

*Perhaps, people reading this book are familiar with naturals, integers, rationals, and reals. However, there are larger sets of numbers as the complex numbers, or the quaternions, the latter introduced by William H. Hamilton already in 1843. Next we provide examples on complex numbers, see their introduction in 1.1.3. The algebra of quaternions is a topic requiring some knowledge from linear algebra. Hence it will be discussed later in paragraph 2.E.66, in terms of the so-called Pauli matrices.<sup>1</sup>*



**1.A.5.** Consider the function  $f(x) = x^2 - 4x + 8$ . Use its graph to ensure yourself that we need to extend the real numbers to solve each quadratic equation with real coefficients. Recall the formula for such a universal solution.

**Solution.** Given a function  $f$  with domain  $X$  and codomain  $Y$ , its graph is the set  $G_f = \{(x, f(x)) : x \in X\}$ . Thus  $G_f$  is a subset of the Cartesian product  $X \times Y$  of  $X, Y$ , i.e.,

$$G_f \subseteq X \times Y = \{(x, y) : x \in X, y \in Y\},$$

where here  $(x, y)$  are ordered pairs, see also 1.6.1.

To illustrate the graph of the given  $f$ , we first note that  $G_f \subseteq \mathbb{R} \times \mathbb{R}$ . This means that  $G_f$  is a subset of the real plane  $\mathbb{R}^2$ , also called *Euclidean plane* (see the discussion in 1.5.1 and 1.5.2 for more details on  $\mathbb{R}^2$ ). To visualize  $G_f$  or parts of it, there are many alternatives. Next we will demonstrate how to use SageMath, which

<sup>1</sup>Complex numbers appear already in *Ars Magna* by Gerolamo Cardano (1501-1576, Italian), around 1545 in relation to solving cubic equations (see 1.A.21 below). Another known number system consists of the *octonions*, also referred to as “Cayley numbers”. Roughly speaking, octonions provide a generalization of quaternions. In a similar fashion, quaternions can be thought of as a generalization of complex numbers, and complex numbers of reals.

William Rowan Hamilton (1805-1865) was an Irish mathematician, astronomer, and physicist. Arthur Cayley (1821-1895) was a British mathematician working mainly in algebra. Actually, the octonions were introduced slightly earlier independently by John T. Graves in 1843, the same year as his friend discovered the quaternions.

Thus the set of natural numbers is either  $\mathbb{Z}^+$ , or the set  $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ . To count “one, two, three, ...” is learned already by children in their pre-school age. Later on, we meet all the integers  $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$  and finally we get used to floating-point numbers. We know what a 1.19-multiple of the price means if we have a 19% tax.

**1.1.1. Properties of numbers.** In order to be able to work properly with numbers, we need to be careful with their definition and properties. In mathematics, the basic statements about properties of objects, whose validity is assumed without the need to prove them, are called *axioms*.



We list the basic properties of the operations of addition and multiplication for our calculations with numbers, which we denote by letters  $a, b, c, \dots$ . Both operations work by taking two numbers  $a, b$ . By applying addition or multiplication we obtain the resulting values  $a + b$  and  $a \cdot b$ .

PROPERTIES OF NUMBERS

**Properties of addition:**

- (CG1)  $(a + b) + c = a + (b + c)$ , for all  $a, b, c$
- (CG2)  $a + b = b + a$ , for all  $a, b$
- (CG3) there exists 0 such that for all  $a$ ,  $a + 0 = a$
- (CG4) for all  $a$  there exists  $b$  such that  $a + b = 0$ .

The properties (CG1)-(CG4) are called the properties of a *commutative group*. They are called respectively *associativity*, *commutativity*, *the existence of a neutral element* (when speaking of addition we usually say zero element), and *the existence of an inverse element* (when speaking of addition we also say the negative of  $a$  and denote it by  $-a$ ).

**Properties of multiplication:**

- (R1)  $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ , for all  $a, b, c$
- (R2)  $a \cdot b = b \cdot a$ , for all  $a, b$
- (R3) there exists 1 such that for all  $a$ ,  $1 \cdot a = a$
- (R4)  $a \cdot (b + c) = a \cdot b + a \cdot c$ , for all  $a, b, c$ .

The properties (R1)-(R4) are called respectively *associativity*, *commutativity*, *the existence of a unit element* and *distributivity of addition with respect to multiplication*. The sets with operation  $+$ ,  $\cdot$  that satisfy the properties (CG1)-(CG2) and (R1)-(R4) are called *commutative rings*. Two further properties of multiplication are:

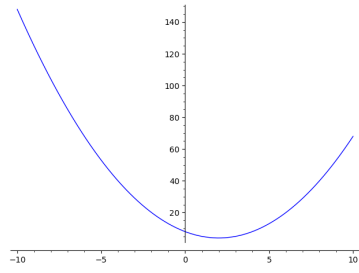
- (F) for every  $a \neq 0$  there exists  $b$  such that  $a \cdot b = 1$ .
- (ID) if  $a \cdot b = 0$ , then either  $a = 0$  or  $b = 0$  or both.

The property (F) is called *the existence of an inverse element* with respect to multiplication (this element is then denoted by  $a^{-1}$ ). For normal arithmetic, this is called the reciprocal of  $a$ , the same as  $1/a$  or  $\frac{1}{a}$ .

The property (ID) then says that there exists no “divisors of zero”. A *divisor of zero* is a number  $a, a \neq 0$ , such that there is a number  $b, b \neq 0$ , with  $a \cdot b = 0$ .

is a free mathematical software system that can be accessed at <https://sagecell.sagemath.org>

The graph of the given function  $f$  is displayed in the image on the right. The Sage cell that generates this graph uses the `plot` command, which requires three parameters: “what to plot”, “what the variable is” and “the range of plotting”. Thus, in general, the syntax we use is as follows:



```
plot(f(x), (x, x_min, x_max))
```

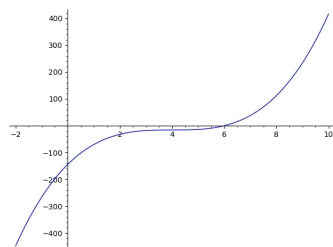
In our case this takes the very simple form

```
f(x) = x**2 - 4*x + 8; plot(f(x), x, -10, 10)
```

In this block observe that we used “;” to type different commands in the same line, while the values  $-10, 10$  specify where  $x$  is evaluated. In the graph we see that  $f$  does not intersect the  $x$ -axis. Thus the equation  $f(x) = x^2 - 4x + 8 = 0$  cannot admit a solution over the reals.

In general, this reflects the negative value of the discriminant  $\Delta = b^2 - 4ac$  of an equation  $ax^2 + bx + c = 0$ . Indeed  $\Delta = -16$  in our case. Recall the well known formula for the solutions  $x_{1,2} = \frac{-b \pm \sqrt{\Delta}}{2a}$ , giving us the (possibly) complex solutions  $x_{1,2}$  once we introduce  $\sqrt{-1} = i$ . Indeed, in our case  $\sqrt{\Delta} = \sqrt{-1}\sqrt{-\Delta} = \sqrt{-1}\sqrt{16} = i\sqrt{16} = 4i$  and the solutions are  $x_{1,2} = 2 \pm i2$ .  $\square$

**1.A.6. A comment on power functions.** Suppose that  $n \in \mathbb{N}$  is a positive natural number.<sup>2</sup> The  $n$ th root  $\sqrt[n]{x} = x^{\frac{1}{n}}$  of some positive real  $x > 0$  is the inverse of the power function  $f(x) = x^n$ , so  $(\sqrt[n]{x})^n = \sqrt[n]{x^n} = x$ . A *polynomial function* is the sum of a finite number of (constant) multiples of power functions with natural exponents, as for example the function  $f(x) = 2(x - 4)^3 - 16$ . Note that it holds  $x^0 = 1$  for each real or complex number  $x$ . Consider the equation  $f(x) = 0$ , i.e.,  $2(x - 4)^3 = 16$ . We have  $(x - 4) = \sqrt[3]{2^3} = 2$ , which gives  $x = 6$  as a quick solution. This is visible also in the graph of  $f$ , given below (this was obtained in Sage via the command `plot(f, -2, 10)`, as before). However, since our equation is of degree 3, over  $\mathbb{C}$  one expects three solutions (based on the fundamental theorem of algebra, see in Chapter 12, and also in the end of this section). In order to determine all the



<sup>2</sup>In this book we mainly adopt the notation  $\mathbb{N} = \{0, 1, 2, 3, \dots\}$  for the (ordered) set of naturals, so we view 0 as a natural number (except Chapter 11 on number theory, where 0 is not considered as a natural number).

**1.1.2. Remarks.** The integers  $\mathbb{Z}$  are a good example of a commutative group. The natural numbers are not such an example since they do not satisfy (CG4) (and possibly do not even contain the neutral element if one does not consider zero to be a natural number). If a commutative ring also satisfies the property (F), we speak of a *field* (often also about a *commutative field*).



The last stated property (ID) is automatically satisfied if (F) holds. However, the converse statement is false. Thus we say that the property (ID) is weaker than (F). For example, the ring of integers  $\mathbb{Z}$  does not satisfy (F) but does satisfy (ID). In such a case we use the term *integral domain*.

Notice that the set of all non-zero elements in the field along with the operation of multiplication satisfies (R1), (R2), (R3), (F) and thus is also a commutative group. However in this case, instead of addition we speak of multiplication. As an example, the set of all non-zero real numbers forms a commutative group under multiplication.

The elements of some set with operations  $+$  and  $\cdot$  satisfying (not necessarily all) stated properties (for example, a commutative field, an integral domain) may be called *scalars*. To denote them we usually use lowercase Latin letters, either from the beginning or from the end of the alphabet.

We will use only these properties of scalars and thus our results will hold for any objects with such properties. This is the true power of mathematical theories – they do not hold just for a specific solved example. Quite the opposite, when we build ideas in a rational way they are always universal. We will try to emphasise this aspect, although our ambitions are modest due to the limited size of this book.

Before coming to any use of scalars, we should make a short formal detour and pay attention to their existence. We shall come back to this in the very end of this chapter, when we shall deal with the formal language of Mathematics in general, cf. the constructions starting in 1.6.5. There we indicate how to get natural numbers  $\mathbb{N}$ , integers  $\mathbb{Z}$ , and rational numbers  $\mathbb{Q}$ , while the real numbers  $\mathbb{R}$  will be treated much later in chapter 5.

At this point, let us just remark that it is not enough to pose the axioms of objects. We have to be sure that the given conditions are not in conflict and such objects might exist.

We suppose the readers are sure about the existence of the domains  $\mathbb{N}$ ,  $\mathbb{Z}$ ,  $\mathbb{Q}$  and can handle them easily. The real numbers are usually understood as a dense and better version of  $\mathbb{Q}$ , but what about the domain of complex numbers?

As is usual in mathematics, we will use variables (letters of alphabet or other symbols) to denote numbers, and it does not matter whether we know their value beforehand or not.

**1.1.3. Complex numbers.** We are forced to extend the domain of real numbers as soon as we want to see solutions of equations like  $x^2 = b$  for all real numbers  $b$ .

We know that this equation always has a solution  $x$  in the domain of real numbers, whenever  $b$  is non-negative. If

solutions, we may use the command `solve` in Sage. This is an appropriate tool for solving equations and system of equations.

In the sequel we will encounter the `solve` command in various contexts and analyze its possible implementations in more detail. For the function  $f$  of our example, we can proceed as follows:

```
x=var("x"); f = 2*(x-4)**3-16
sol = solve([f==0], x); sol
```

and as an output we obtain a list of the three solutions, with two of them being complex conjugate, i.e.,

```
[x == -I*sqrt(3) + 3,
 x == I*sqrt(3) + 3, x == 6]
```

A traditional way to find the roots of polynomials relies on the so called *Horner's scheme*. This provides a quick way to evaluate a given polynomial  $f(x)$ , and occurs by rewriting the polynomial as follows:

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots + a_nx^n = a_0 + x(a_1 + x(a_2 + x(a_3 + \dots + x(a_{n-1} + xa_n) \dots)))$$

Observe that such a procedure requires  $n$  multiplications and  $n$  additions (and this turns out to be optimal). In fact we can treat the division of polynomials by divisors of the form  $x - \rho$  ( $\rho \in \mathbb{R}$ ), using the *Horner's table*:

$\rho$	$a_n$	$a_{n-1}$	$\dots$	$a_0$
	-	$\rho a_n$	$\dots$	
	$a_n$	$\rho a_n + a_{n-1}$	$\dots$	$f(\rho)$

In this table, each entry in the second row is the product of  $\rho$  with the bottom-row entry immediately to the left. The bottom line is the sum of the previous two lines. For example, if  $f(x) = x^3 + x^2 - x + 2$  and  $\rho = -2$ , we obtain

$\rho = -2$	1	1	-1	2
		-2	2	-2
	1	-1	1	0

The zero in the last entry (the remainder), verifies that  $\rho = -2$  is a root of our polynomial, i.e.,  $f(\rho) = 0$ .<sup>3</sup>

**I.A.7.** Use the Horner's scheme to find the division of  $p(x) = 5x^4 + x^2 - 2x + 2$  by  $x - 4$ . Then verify your answer by the classical algorithm of (long) division. ○

Polynomials can be defined also over the complex numbers, and the Horner's scheme is applicable in this case as well. Recall that to each complex number  $z = x + iy \in \mathbb{C}$  we can associate the ordered pair  $(x, y)$  in the Euclidean plane. Thus, complex numbers can be viewed as ordered pairs of real numbers  $(x, y) \in \mathbb{R}^2$ , and such pairs correspond to vectors on  $\mathbb{R}^2$  with initial point the origin of  $\mathbb{R}^2$ . As usual we will use the  $x$ -axis for the reals  $\text{Re } z$ , and the  $y$ -axis to denote the purely



<sup>3</sup>The motivation back of the notation  $\rho$  in the Horner's scheme, is due to the greek word " $\rho\iota\chi\alpha$ ", which means "root".

$b < 0$ , then such a real  $x$  cannot exist. Thus we need to find a larger domain, where this equation has a solution.

The crucial idea is to add the new number  $i$  to the real numbers, the *imaginary unit*, for which we require  $i^2 = -1$ . Next we try to extend the definitions of addition and multiplication in order to preserve the usual behaviour of numbers (as summarised in 1.1.1).

Clearly we need to be able to multiply the new number  $i$  by real numbers and sum it with real numbers. Therefore we need to work in our newly defined domain of *complex numbers*  $\mathbb{C}$  with formal expressions of the form  $z = a + ib$ , being called *algebraic form* of  $z$ . The real number  $a$  is called the *real part* of the complex number  $z$ , the real number  $b$  is called the *imaginary part* of the complex number  $z$ , and we write  $\text{Re}(z) = a$ ,  $\text{Im}(z) = b$ . It should be noted that if  $z = a + ib$  and  $w = c + id$  then  $z = w$  implies both  $a = c$  and  $b = d$ . In other words, we can equate both real and imaginary parts. For positive  $x$  we then get  $(i \cdot x)^2 = -1 \cdot x^2$  and thus we can solve the equations as requested.

In order to satisfy all the properties of associativity and distributivity, we define the addition so that we add independently the real parts and the imaginary parts. Similarly, we want the multiplication to behave as if we multiply the pairs of real numbers, with the additional rule that  $i^2 = -1$ , thus

$$(a + ib) + (c + id) = (a + c) + i(b + d),$$

$$(a + ib) \cdot (c + id) = (ac - bd) + i(bc + ad).$$

Next, we have to verify all the properties (CG1-4), (R1-4) and (F) of scalars from 1.1.1. But this is an easy exercise: zero is the number  $0 + i0$ , one is the number  $1 + i0$ , both these numbers are for simplicity denoted as before, that is, 0 and 1. For non-zero  $z = a + ib$  we easily check that  $z^{-1} = (a^2 + b^2)^{-1}(a - ib)$ . All other properties are obtained by direct calculations.

**1.1.4. The complex plane and polar form.**

A complex number is given by a pair of real numbers, therefore it corresponds to a point in the real plane  $\mathbb{R}^2$ . Our algebraic form of the complex numbers  $z = x + iy$  corresponds in this picture to understanding the  $x$ -coordinate axis as the real part while the  $y$ -coordinate axis is the imaginary part of the number. The *absolute value* of the complex number  $z$  is defined as its distance from the origin, thus  $|z| = \sqrt{x^2 + y^2}$ .

The reflection with respect to the real axis then corresponds to changing the sign of the imaginary part. We call this operation  $z \mapsto \bar{z} = x - iy$  the *complex conjugation*.

Let us now consider complex numbers of the form  $z = \cos \varphi + i \sin \varphi$ , where  $\varphi$  is a real parameter giving the angle between the real axis and the line from the origin to  $z$  (measured in the positive, i.e. counter-clockwise sense). These numbers describe all points on the unit circle in the complex plane. Every non-zero complex number  $z$  can be then written as

$$z = |z|(\cos \varphi + i \sin \varphi).$$

imaginary numbers  $i \operatorname{Im} z$ . We proceed with basic exercises on the arithmetic of complex numbers. Recall that both the addition and the multiplication of complex numbers attain a geometric interpretation: Addition corresponds to vector sum, while multiplication by  $i = \sqrt{-1}$ , for example, is equivalent to a counterclockwise rotation by the right angle.

**1.A.8. Arithmetic in  $\mathbb{C}$ .** Compute  $\operatorname{Re}(z_1)$ ,  $\operatorname{Im}(z_1)$ ,  $\bar{z}_1$ ,  $|z_2|$ ,  $z_1 + z_2$ , and  $z_1 z_2$ , for the cases:

- (a)  $z_1 = 1 - 2i$ ,  $z_2 = 4i - 3$ ;
- (b)  $z_1 = 3 + 4i$ ,  $z_2 = 3 - 4i$ .

**Solution.** (a) We have  $\operatorname{Re}(z_1) = 1$ ,  $\operatorname{Im}(z_1) = -2$ ,  $\bar{z}_1 = 1 + 2i$ ,  $|z_2| = \sqrt{4^2 + (-3)^2} = 5$ ,  $z_1 + z_2 = -2 + 2i$ ,  $z_1 z_2 = 5 + 10i$ .

(b) Here we use Sage, where the field of complex numbers is denoted by `CC`, while as we saw above the square root of  $-1$  is denoted by `I` (or `i`). To introduce a complex number there are many alternatives, e.g., we simply type `z1 = 3 + 4 * I`, where `*` is the general multiplication operator in Sage<sup>4</sup>, or type `z1 = CC(3, 4)`. As for the additional code that gives the computations, type and run successively the following

```
z1 = CC(3, 4); z2 = CC(3, -4)
z1.real (); z1.imag ()
z1.conjugate (); abs(z2); z1+z2; z1*z2
```

Note that in case (b) we have  $z_2 = \bar{z}_1$ . In Sage a verification of this takes the form `z1 == conjugate(z2)`, after introducing  $z_1, z_2$  as above. As it is usual also in Python (the mother of Sage), this means that one uses “=” to determine a quantity, instead of “==” which is used to determine an equation between two quantities (in the latter case Sage returns `True` or `False`). It follows that  $|z_1| = |z_2| = \sqrt{25} = 5$  (see also 1.A.10 below), and the multiplication  $z_1 z_2$  occurs also by the usual formula  $z\bar{z} = |z|^2$ , with  $z \in \mathbb{C}$  (verify this).  $\square$

**1.A.9.** Show that the Euclidean distance between two arbitrary complex numbers  $z = x + iy$  and  $w = u + iv$  is given by the formula  $d = |z - w| = \sqrt{(x - u)^2 + (y - v)^2}$ , with  $x, y, u, v \in \mathbb{R}$ .  $\circ$

**1.A.10.** Given the complex number  $1 + i\sqrt{3}$ , make use only of the figure given below to calculate its distance from its complex conjugate. Next, consider the complex number  $z = \frac{2+2i\sqrt{3}+3i-3\sqrt{3}}{1-i\sqrt{3}}$ . Relate a verification of the equality  $|z| = \sqrt{13}$  with the given figure and find another  $z' \in \mathbb{C}$  with the same property, i.e.,  $|z'| = \sqrt{13}$ .

**Solution.** Complex conjugates are symmetric with respect to the  $x$ -axis. Moreover, the distance of a complex number from the  $x$ -axis equals its imaginary part. Thus, as we deduce also from the given figure, the distance between  $1 + i\sqrt{3}$  and its conjugate  $1 - i\sqrt{3}$  equals to  $2\sqrt{3}$ . Now, from the Pythagorean

<sup>4</sup>As we will see later, in Sage the multiplication of matrices and vectors is also denoted by `*`.

For given  $z \neq 0$ ,  $\varphi$  is unique if  $0 \leq \varphi < 2\pi$ . The number  $\varphi$  is called the argument of the complex number  $z$  and this form of  $z$  is called the *polar form* of the complex number. This way of writing the complex numbers is very convenient for understanding the multiplication.

Consider the numbers  $z = |z|(\cos \varphi + i \sin \varphi)$  and  $w = |w|(\cos \psi + i \sin \psi)$  and calculate their product

$$\begin{aligned} z \cdot w &= |z|(\cos \varphi + i \sin \varphi)|w|(\cos \psi + i \sin \psi) \\ &= |z||w|(\cos \varphi \cos \psi - \sin \varphi \sin \psi \\ &\quad + i(\cos \varphi \sin \psi + \sin \varphi \cos \psi)) \\ &= |z||w|(\cos(\varphi + \psi) + i \sin(\varphi + \psi)). \end{aligned}$$

The last equality is a result of the addition formulas for trigonometric functions (we shall deal with them in more detail later in our discussion of rotations in the plane, see the page 37).

Division is equally easy. If  $z = |z|(\cos \varphi + i \sin \varphi) \neq 0$ , then  $w = |z|^{-1}(\cos \varphi - i \sin \varphi)$  satisfies  $zw = wz = 1$ , hence we can write  $w = z^{-1} = 1/z$ .

We can summarize (and iterate the application of the previous formula on the product of the number  $z$  with itself):

POLAR FORM AND DE MOIVRE THEOREM

Consider two complex numbers  $z = |z|(\cos \varphi + i \sin \varphi)$  and  $w = |w|(\cos \psi + i \sin \psi)$  in polar forms. Then if  $n$  is an integer, positive or negative,

$$\begin{aligned} z w &= |z||w|(\cos(\varphi + \psi) + i \sin(\varphi + \psi)) \\ z^n &= |z|^n(\cos(n\varphi) + i \sin(n\varphi)). \end{aligned}$$

**1.1.5. Functions.** In most tasks we do not deal just with numbers, i.e. with individual values of scalars. More often the values are associated to each of the elements in a set of objects.



Formally we talk about a *mapping*  $f : A \rightarrow B$  assigning to each element  $x$  in the *domain* set  $A$  the value  $f(x)$  in the *codomain* set  $B$ . The set of all images  $f(x) \in B$  is called the *range* of  $f$ .

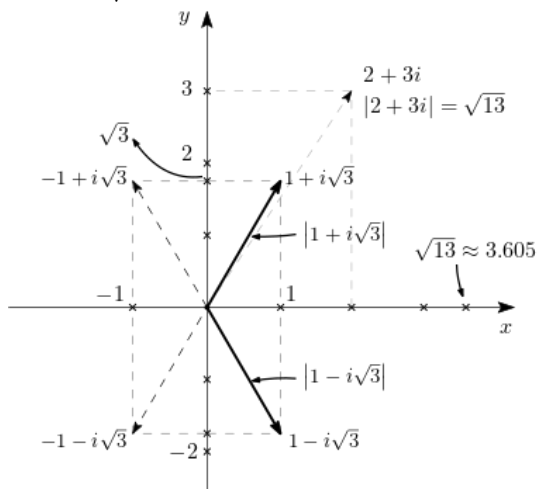
The set  $A$  or  $B$  can be a set of numbers, but there is nothing to stop them being sets of other objects. The mapping  $f$ , however it is described, must unambiguously determine a unique member of  $B$  for each member of  $A$ .

In another terminology, the member  $x \in A$ , is often called the *independent variable*. Then  $y = f(x) \in B$ , is called the *dependent variable*. We also say that the value  $y = f(x)$  is a *function* of the independent variable  $x$  in the domain of  $f$ .

For now, we shall restrict ourselves to the case where the codomain  $B$  is a subset of scalars and we shall talk about *scalar functions*.

The simplest way to define a function appears if  $A$  is a finite set. Then we can describe the function  $f$  by a table or a listing showing the image of each member of  $A$ . We have certainly seen many examples of such functions:

Theorem it follows that a complex number  $x + iy$  and its complex conjugate  $x - iy$  have the same length (absolute value), given by  $r := \sqrt{x^2 + y^2}$ .



For our task this applies to  $1 + i\sqrt{3}$ . On the other hand, recall that the absolute value of the product of any two complex numbers is the product of their absolute values. Thus, if  $z_1, z_2$  are two complex numbers, provided that  $z_2 \neq 0$ , we see that

$$\left| \frac{z_1 z_2}{z_2} \right| = \frac{|z_1 z_2|}{|z_2|} = \frac{|z_1| |z_2|}{|z_2|} = |z_1|.$$

Observe now the numerator of the given  $z$  is expressed as

$$\begin{aligned} 2 + 2i\sqrt{3} + 3i - 3\sqrt{3} &= 2(1 + i\sqrt{3}) + 3i(1 + i\sqrt{3}) \\ &= (2 + 3i)(1 + i\sqrt{3}). \end{aligned}$$

This means that our  $z$  has the form  $\frac{z_1 z_2}{z_2}$  with  $z_1 = 2 + i3$  and  $z_2 = 1 + i\sqrt{3}$ , respectively, and hence the result:  $|z| = |z_1| = |2 + i3| = \sqrt{2^2 + 3^2} = \sqrt{13}$ . Because  $z_2$  does not play any essential role in this computation, one can find infinitely many complex numbers with the same property as the given  $z$ .  $\square$

**1.A.11.** Determine the distance  $d$  of the numbers  $z, \bar{z}$  in the complex plane for  $z = \sqrt{\sqrt{13}} - i(\sqrt{13}/2)$ .  $\circ$

**1.A.12.** Use the Horner's scheme to solve the equation

$$z^3 + (1 + i)z^2 + az + 2 = 0, \quad z \in \mathbb{C},$$

when it is already known that  $z_0 = i$  is a root of the given equation.  $\circ$

**1.A.13.** Solve the equation  $x^2 + 2(1 + i)x = 14i$  (observe that this is a quadratic equation with complex coefficients).

**Solution.** Add to both sides of the equation the term  $(1 + i)^2$ . This gives  $x^2 + 2(1 + i)x + (1 + i)^2 = 14i + (1 + i)^2$ , or equivalently,

$$(x + (1 + i))^2 = 16i.$$

Therefore,  $x + (1 + i) = \pm 4\sqrt{i}$ , that is,  $x = -(1 + i) \pm 4\sqrt{i}$  and it remains to compute  $\sqrt{i}$ . Suppose that  $a + ib = \sqrt{i}$  for some reals  $a, b \in \mathbb{R}$ . Then  $(a + ib)^2 = i$  or in other words  $a^2 + 2iab - b^2 = i = 0 + 1i$ . By comparing the real and imaginary parts, we obtain the equations  $a^2 - b^2 = 0$  and  $2ab = 1$ . The first relation gives  $a = \pm b$ . If  $a = b$  then the

Let  $f$  denote the pay of a worker in some company in certain year. The values of independent variable, that is, the domain of the function, are individual workers  $x$  from the set of all considered workers. The value  $f(x)$  is their pay for the given year. Similarly we can talk about the age of students or their teachers in years, the litres of beer and wine consumed by individuals from a given group, etc.

Another example is a food dispensing machine. The domain of a function  $f$  would be the button pushed together with the money inserted to determine the selection of the food item.

Let  $A = \{1, 2, 3\} = B$ . The set of equalities  $f(1) = 1, f(2) = 3, f(3) = 3$ , defines a function  $f : A \rightarrow B$ . Generally, as there are 3 possible values for  $f(1)$ , and the same for  $f(2)$ , and  $f(3)$ , there are 27 possible functions from  $A$  into  $B$  in total.

But there are other ways to define a function than as a table. For example, the function  $f$  can denote the area of a planar region. Here, the domain consists of subsets of the plane (e.g. all triangles, circles or other planar regions with a defined area). The range of  $f$  consists of the respective areas of the regions. Rather than providing a list of areas for a finite number regions, we hope for a formula allowing us to compute the functional value  $f(P)$  for any given planar region  $P$  from a suitable class.

Of course, there are many simple functions given by formulae, e.g.  $f(x) = 3x + 7$  with  $A = B = \mathbb{R}$  or  $A = B = \mathbb{N}$ .

Not all functions can be given by a formula or list. For example, let  $f(t)$  denote the speed of the car at time  $t$ . For any given car and time  $t$ , we know there will be the functional values  $f(t)$  denoting its speed. Which can of course be measured approximately, but usually not by a formula.

Another example: Let  $f(n)$  be the  $n^{\text{th}}$  digit in the decimal expansion of  $\pi \doteq 3.1415\dots$ . So for example  $f(4) = 5$ . The value of  $f(n)$  is defined but unknown if  $n$  is large enough.

The mathematical approach in modelling real problems often starts from the indication of certain dependencies between some quantities and aims at explicit formulas for functions which describe them. Often a full formula is not available but we may obtain the values  $f(x)$  at least for some instances of the independent variable  $x$ , or we may be able to find a suitable approximation.

We shall see all of the following types of expressions of the requested function  $f$  in this book:

- exact finite expression (like the function  $f(x) = 3x + 7$  above);
- infinite expression (we shall come to that only much later in chapter 5 when introducing the limit processes);
- description of how the function's values change under a given change of the independent variable (this behaviour will be displayed under the name difference equation in a moment and under different circumstances later on);
- approximation of a not computable function with a known one (usually including some error estimates – this could be the case with the car above, say we know

second relation reduces to  $a^2 = 1/2$ , that is,  $a = \pm 1/\sqrt{2} = \pm \sqrt{2}/2 = b$ . We also see that the case  $a = -b$  is impossible; It gives  $2a^2 = -1$  which contradicts to the fact  $a \in \mathbb{R}$ . Thus  $\sqrt{i} = \pm \frac{\sqrt{2}}{2}(1 + i)$  and

$$x = -(1 + i) \pm 2\sqrt{2}(1 + i) = (-1 \pm 2\sqrt{2})(1 + i).$$

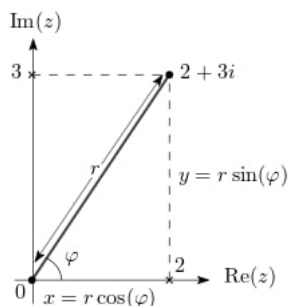
For convenience, let us verify the answer via Sage. The next commands successively written, yield the computations of the real/imaginary parts of the two complex solutions:

```
eq = x**2+2*(1+I)*x-14*I==0
sols = solve([eq], x); sols
sols[0].rhs().real_part()
sols[0].rhs().imag_part()
sols[1].rhs().imag_part()
sols[1].rhs().real_part()
```

For example, the last command returns  $2 * \text{sqrt}(2) - 1$ , as it should be, and similarly for the previous commands. As a useful remark, keep in mind that in Sage the command `sols[0]` returns the first solution, and `sols[1]` the second one. This means that Sage assigns automatically labels to these solutions, as the first two integers “0”, and “1”, respectively.  $\square$

**1.A.14. Polar form and algebraic form.** Express  $z_1 = 2 + 3i$  in polar form. Next, express  $z_2 = 3(\cos(\pi/3) + i \sin(\pi/3))$  in algebraic form.

**Solution.** As we learned from the figure in 1.A.10, the absolute value  $|z_1|$  of the complex number  $z_1$  equals to  $r = \sqrt{13}$ . Next, we need the angle (argument)  $\varphi$  of  $z_1$ , indicated in the figure. Applying the well-known rules for the trigonometric functions  $\cos$  and  $\sin$  yields  $\sin(\varphi) = 3/\sqrt{13}$  and  $\cos(\varphi) = 2/\sqrt{13}$ , respectively. Therefore, with the aid of a calculator we get



$$\varphi = \arcsin(3/\sqrt{13}) = \arccos(2/\sqrt{13}) \doteq 56.3^\circ.$$

This computation is also easy in Sage: Running the cell

```
z=2+3*I; arg(z)
```

we obtain  $\arctan(3/2)$ . Then, to translate the angle in radians or degrees we just type

```
phi = arctan(3/2)
print(N(phi,digits = 5), "radians")
print(N(phi*180/pi,digits=4), "degrees")
```

The given answer is 0.98279 radians and 56.31 degrees, respectively. Now we are able to present the polar form of  $z_1 = x + iy$ : We have  $x = \sqrt{13} \cos(\varphi)$  and  $y = \sqrt{13} \sin(\varphi)$

it goes with some known speed at the time  $t = 0$ , we break as much as possible on a known surface and we compute the decrease of speed with the help of some mathematical model);

- finding only the probability of possible values of the function. For example the function giving the length of life of a given group of still living people, in dependence of some health related parameters.

**1.1.6. Functions defined explicitly.** Let us start with the



most desirable case, when the function values are defined by a computable finite formula. Of course, we shall be interested also in the efficiency of the formulas, i.e. how fast the evaluations would be. In principle, real computations can involve only a finite number of summations and multiplications of numbers. This is how we define the *polynomials*, i.e. functions of the form  $f(x) = a_n \cdot x^n + \dots + a_1 \cdot x + a_0$ , where  $a_0, \dots, a_n$  are known scalars,  $x$  is the unknown variable whose value we can insert.  $x^n = 1 \cdot x \cdot \dots \cdot x$  means the  $n$ -times repeated multiplication of the unit by  $x$  (in particular,  $x^0 = 1$ ), and  $f(x)$  is the value of the indicated sum of products. This is fairly well computable formula for each  $n \in \mathbb{N}$ . The choice  $n = 0$  provides the constants (i.e. constant functions)  $a_0$ .

The next example is more complicated.

### FACTORIAL FUNCTION

Let  $A = \mathbb{Z}^+$  be the set of positive integers. For each  $n \in \mathbb{Z}^+$ , define the factorial function by

$$n! = n(n-1)(n-2) \dots 3 \cdot 2 \cdot 1.$$

For convenience we also define  $0! = 1$ . (We will see why this is sensible later on). It is easy to see that  $n! = n \cdot (n-1)!$  for all  $n \geq 1$ .

So  $1! = 1$ ,  $2! = 2 \cdot 1 = 2$ ,  $3! = 3 \cdot 2 \cdot 1 = 6$ ,  $6! = 720$  etc.

The latter example deserves more attention. Notice that we could have defined the factorial by setting  $A = B = \mathbb{N}$  and giving the equation  $f(n) = n \cdot f(n-1)$  for all  $n \geq 1$ . This does not yet define  $f$ , but for each  $n$ , it does determine what  $f(n)$  is in terms of its predecessor  $f(n-1)$ . This is sometimes called a *recurrence relation*. After choosing  $f(0) = 1$ , the recurrence now determines  $f(1)$  and hence successively  $f(2)$  etc., and so a function is defined. It is the factorial function as described above.

## 2. Difference equations



The factorial function is one example of a function which can be defined on the natural numbers by means of a recurrence relation. Such a situation can often be seen when formulating mathematical models that describe real systems in practice.

We will observe here only a few simple examples and return to this topic in chapter 3.

which gives

$$\begin{aligned} z_1 &= \sqrt{13} \left( \frac{2}{\sqrt{13}} + i \cdot \frac{3}{\sqrt{13}} \right) \\ &= \sqrt{13} \left( \cos \left( \arccos \left( \frac{2}{\sqrt{13}} \right) \right) + i \sin \left( \arcsin \left( \frac{3}{\sqrt{13}} \right) \right) \right). \end{aligned}$$

Transition from polar form to algebraic form is even simpler:

$$z_2 = 3 \left( \cos \left( \frac{\pi}{3} \right) + i \sin \left( \frac{\pi}{3} \right) \right) = 3 \left( \frac{1}{2} + i \cdot \frac{\sqrt{3}}{2} \right). \quad \square$$

**1.A.15. Remark.** In principle, the argument  $\arg(z) = \varphi$  of a complex number  $z$  is only defined up to “modulo  $2\pi$ ”. That is, for any integer  $k \in \mathbb{Z}$ , the angle  $\varphi + 2k\pi$  would serve as well (since full circle rotations do not change the point). Therefore, in the previous exercise we have essentially defined the so-called *principal (value of the) argument* of  $z$ .

**1.A.16.** (a) Is it true that the (principal) argument of a positive real number is zero and of a negative one is  $\pi$ ?

(b) Is it true that the (principal) argument of the imaginary number  $i$  is  $\pi/2$ , while of  $-2i$  is  $-\pi$ ? ○

**1.A.17.** Express  $z = 1 + \cos \frac{\pi}{3} + i \sin \frac{\pi}{3}$  in polar form. ○

**1.A.18. De Moivre’s theorem applied.** Simplify the expression  $(5\sqrt{3} + 5i)^n$  for  $n = 2$  and  $n = 12$ .

**Solution.** Using the binomial theorem for  $n = 2$  we see that

$$(5\sqrt{3} + 5i)^2 = 75 + 10\sqrt{3} \cdot 5i - 25 = 50 + 50\sqrt{3}i.$$

For  $n = 12$  it is shorter (and wiser) to express first the complex number in polar form:

$$5\sqrt{3} + 5i = 10 \left( \frac{\sqrt{3}}{2} + \frac{i}{2} \right) = 10 \left( \cos \frac{\pi}{6} + i \sin \frac{\pi}{6} \right).$$

Then, an application of the de Moivre’s formula presented in 1.1.4, gives the following:

$$(5\sqrt{3} + 5i)^{12} = 10^{12} \left( \cos \frac{12\pi}{6} + i \sin \frac{12\pi}{6} \right) = 10^{12}. \quad \square$$

**1.A.19.** Compute the expression  $(\cos \frac{\pi}{6} + i \sin \frac{\pi}{6})^{31}$  by applying the de Moivre’s theorem. ○

**1.A.20.** Use de Moivre’s theorem to prove the identities

$$\begin{aligned} \cos(3\varphi) &= 4 \cos^3(\varphi) - 3 \cos(\varphi), \\ \sin(3\varphi) &= 3 \sin(\varphi) - 4 \sin^3(\varphi). \end{aligned} \quad \circ$$

The beauty of complex numbers is also linked to the “fundamental theorem of algebra”, see 12.4.20. In fact, complex numbers arose from the need to solve cubic equations (without meaning that the solutions must be complex). Next we describe a procedure, which was developed during the 16th century by S. del Ferro (1465-1526), G. Cardano (1501-1576), N. F. Tartaglia (1500-1557), and possibly others. Unlike the



**1.2.1. Linear difference equations of first order.** A general *difference equation of the first order* (or *first order recurrence*) is an expression of the form

$$f(n+1) = F(n, f(n)),$$

where  $F$  is a known function with two arguments (independent variables). If we know the “initial” value  $f(0)$ , we can compute  $f(1) = F(0, f(0))$ , then  $f(2) = F(1, f(1))$  and so on. Using this, we can compute the value  $f(n)$  for arbitrary  $n \in \mathbb{N}$ .

An example of such an equation is provided by the factorial function  $f(n) = n!$  where:

$$(n+1)! = (n+1) \cdot n!, \quad f(0) = 1.$$

In this way, the value of  $f(n+1)$  depends on both  $n$  and the value of  $f(n)$ , and formally we would express this recurrence in the form  $F(x, y) = (x+1)y$ .

A very simple example is  $f(n) = C$  for some fixed scalar  $C$  and all  $n$ . Another example is the *linear difference equation of first order*

$$(1) \quad f(n+1) = a \cdot f(n) + b,$$

where  $a \neq 0$  and  $b$  are fixed scalars.

Such a difference equation is easy to solve if  $b = 0$ . Then it is the well-known recurrent definition of the geometric progression. We have

$$f(1) = af(0), \quad f(2) = af(1) = a^2f(0),$$

and so on. Hence for all  $n$  we have

$$f(n) = a^n f(0).$$

This is also the relation for the *Malthusian population growth model*. This is based on the assumption that population size grows with a constant rate when measured at a sequence of fixed time intervals.

It is time to prove our first mathematical theorem. We deduce a general result for first order equations with variable coefficients, namely:

$$(2) \quad f(n+1) = a_n \cdot f(n) + b_n.$$

We use the usual notation for sum  $\sum$ , and the similar notation for the product  $\prod$ . We use also the convention that when the index set is empty, then the sum is zero and the product is one.

**1.2.2. Proposition.** *The general solution of the first order difference equation (2) from the previous paragraph with the initial condition  $f(0) = y_0$  is for  $n \in \mathbb{N}$  given by the formula*

$$(1) \quad f(n) = \left( \prod_{i=0}^{n-1} a_i \right) y_0 + \sum_{j=0}^{n-2} \left( \prod_{i=j+1}^{n-1} a_i \right) b_j + b_{n-1}.$$

**PROOF.** We use *mathematical induction*. The result clearly holds for  $n = 1$  since  $f(1) = a_0 y_0 + b_0$ .

Assuming that the statement holds for some fixed  $n$ , we compute (do not be upset with the many brackets and remember the conventions about the empty sums and products):





quadratic equations, here we need the complex numbers even if all three solutions are real.

**1.A.21.** Express a solution of the cubic equation  $x^3 + ax^2 + bx + c = 0$  in terms of the real coefficients  $a, b$  and  $c$ .

**Solution.** The first step reduces the quadratic dependence by setting  $x := t - a/3$ . The original equation then becomes  $t^3 + pt + q = 0$ , with  $p = b - a^2/3$  and  $q = c + (2a^3 - 9ab)/27$ , respectively. The next step requires us to be more creative. We introduce new variables,  $u$  and  $v$ , satisfying the conditions

$$\{u + v = t, \quad 3uv + p = 0\}.$$

Using these terms, we can substitute the first condition into the previous equation to obtain

$$u^3 + v^3 + (3uv + p)(u + v) + q = 0.$$

Next, use the second equation to eliminate  $v$ . This yields

$$u^6 + qu^3 - \frac{p^3}{27} = 0,$$

which is a quadratic equation in the unknown  $s = u^3$ . Hence

$$u = \sqrt[3]{-\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}.$$

Finally, by back substitution we obtain the desired answer:

$$x = -p/3u + u - a/3.$$

Observe that in order to obtain all three solutions, one has to work with complex roots. This is because the equation  $x^3 = a$ , with  $a \neq 0$ , has exactly three solutions over  $\mathbb{C}$  (according to the fundamental theorem of algebra). If one needs all the three solutions, the Sage cell given below provides an answer (though the expressions are complicated enough and hence not presented here).

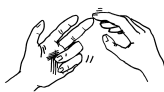
```
x, a, b, c=var("x, a, b, c")
assume(a, b, c, "real")
solve(x**3+a*x**2+b*x+c==0, x) □
```

**1.A.22.** Solve the equation  $x^3 + x^2 - 2x - 1 = 0$ . ○

*A series of additional tasks related to complex numbers is presented in the end of this chapter. In fact, once we provide a systematic treatment of basic linear algebra in Chapter 2, we will be able to demonstrate how to treat complex numbers via matrices. This takes place in paragraph ??.*

### B. Difference equations

*Difference equations (also called recurrence relations) are rules which determine the values of elements of a sequence in terms of previous elements. Solving a difference equation means finding an explicit formula for an arbitrary element of the sequence. If each element of the sequence is determined only by the previous element, we talk about first order difference equations. As we will see below, and also in Chapter 3, this*



$$\begin{aligned} f(n+1) &= a_n \left( \prod_{i=0}^{n-1} a_i \right) y_0 + \sum_{j=0}^{n-2} \left( \prod_{i=j+1}^{n-1} a_i \right) b_j + b_{n-1} \\ &\quad + b_n \\ &= \left( \prod_{i=0}^n a_i \right) y_0 + \sum_{j=0}^{n-1} \left( \prod_{i=j+1}^n a_i \right) b_j + b_n, \end{aligned}$$

as can be seen directly by multiplying out. □

Note that for the proof, we did not use anything about the scalars except for the properties of commutative ring.

There is another approach to the proof which explains the name "linear". First observe that with all  $b_n = 0$ , the formula from the proposition is obvious. Next observe, that dealing with two sequences  $y_n, y'_n$  solving 1.2.2(2), their difference  $z_n = y_n - y'_n$  must be a solution to the "homogenized" equation  $f(n+1) = a_n \cdot f(n)$ . On the contrary, knowing such sequences  $z_n$  and  $y_n$ , then their sum  $y'_n = z_n + y_n$  must be a solution of 1.2.2(2). Thus, we could just guess and check that the second and third terms of (1) solve our equation. In particular, the following corollary can be proved very easily this way. Try to complete this line of arguments in detail!

**1.2.3. Corollary.** *The general solution of the linear difference equation (1) from 1.2.1 with  $a \neq 1$  and initial condition  $f(0) = y_0$  is*

$$(1) \quad f(n) = a^n y_0 + (1 + \dots + a^{n-1})b = a^n y_0 + \frac{1 - a^n}{1 - a} b.$$

**PROOF.** If we set  $a_i$  and  $b_i$  to be constants and use the general formula 1.2.2(1), we obtain

$$f(n) = a^n y_0 + b \left( 1 + \sum_{j=0}^{n-2} a^{n-j-1} \right).$$

We observe that the expression in the bracket is  $(1 + a + \dots + a^{n-1})$ . The sum of this geometric progression follows from  $1 - a^n = (1 - a)(1 + a + \dots + a^{n-1})$ . □

The proof of the former proposition is a good example of a mathematical result, where the verification is quite easy, as soon as someone tells us the theorem. Mathematical induction is a natural method of proof.

Note that for calculating the sum of a geometric progression we silently assumed the existence of the inverse element for non-zero scalars. But actually, we derived the formula in a way manifesting that the division is always possible if our scalars do not allow for divisors of zero.

In particular, notice that the formula (1) is valid with integer coefficients  $a, b$  and integer initial conditions. Here, we know in advance that each  $f(n)$  is an integer. Thus our formula necessarily gives correct integer solutions, although living in the extension of  $\mathbb{Z}$  to the rational numbers  $\mathbb{Q}$ .

type of recurrence relations are often induced by problems appearing in our everyday life.

**1.B.1.** Michael wants to buy a great new car, which costs € 30 000. He wants to take out a loan and pay it off under a fixed month repayment agreement. The car company offers him a loan with yearly interest of 6%, with first repayment in the end of the first month of the loan. Specify the exact amount of the monthly payment, under the assumption that Michael aims to redeem the loan in three years.

**Solution.** Let us denote by  $P$  the sum that Michael has to pay per month, and by  $d_k$  the amount of the remaining loan after  $k$  months. Set also  $C = 30000$  and  $u = \frac{0.06}{12}$  (the latter represents the monthly interest rate). We have  $d_0 = C = 30000$ , and after the first month we see that  $d_1 = C - P + u \cdot C$ . After the  $k$ th month we compute

$$d_k = d_{k-1} - P + u d_{k-1} = (1 + u)d_{k-1} - P. \quad (*)$$

In order to solve (\*) we may apply the statement of Corollary 1.2.3. In terms of this result we have  $a = 1 + u$ ,  $b = -P$ , and thus the solution takes the form

$$d_k = d_0 a^k - P \left( \frac{a^k - 1}{a - 1} \right).$$

Paying off the loan in three years means  $d_{36} = 0$ . This gives

$$P = 30000 \cdot \left( \frac{(1 + u)^{36} u}{(1 + u)^{36} - 1} \right) = 912,7.$$

Thus, approximately Michael should pay the amount of €913 per month.  $\square$

**1.B.2.** Consider the task in 1.B.1. For how long will Michael pay, if the monthly dose of the loan is €500?  $\circ$

**1.B.3.** Determine the sequence  $\{y_n\}_{n=1}^\infty$ , which satisfies the following recurrence relation

$$y_{n+1} = \frac{3y_n}{2} + 1, \quad n \geq 1, \quad y_1 = 1.$$

Next verify your answer via Sage.

**Solution.** We can apply again the Corollary 1.2.3. We have  $a = 3/2$ ,  $b = 1$  and  $y_0 = 0$  (such that  $y_1 = 1$ ). Hence, according to the given formula we deduce that

$$y_n = (3/2)^n \cdot 0 + \frac{1 - (3/2)^n}{1 - \frac{3}{2}} \cdot 1 = -2 + 2(3/2)^n.$$

Let us explain the procedure via Sage, which has a very friendly package to treat recurrence relations. An appropriate function that one can load is the function `rsolve` (from the pure Python package `sympy` which is for symbolic computations). This can deal with linear recurrence relations and initially we type the following code:

```
from sympy import Function, rsolve
from sympy.abc import n
y = Function("y")
```

The next step is to type the relation that we want to solve and the initial conditions:



The linear difference equation 1.2.1(1) can be neatly interpreted as a mathematical model for finance, e.g. savings or loan payoff with a fixed interest rate  $a$  and fixed repayment  $b$ . (The cases of savings and loans differ only in the sign of  $b$ ).

With varying parameters  $a$  and  $b$  we obtain a similar model with varying interest rate and repayment. We can imagine for instance that  $n$  is the number of months,  $a_n$  is the interest rate in the  $n$ th month,  $b_n$  the repayment in the  $n$ th month.

**1.2.4. A nonlinear example.** When discussing linear difference equations, we mentioned a very primitive population growth model which depends directly on the momentary population size  $p$ . At first sight, it is clear that such a model with  $a > 1$  leads to a very rapid and unbounded growth.



A more realistic model has such a population change  $\Delta p(n) = p(n + 1) - p(n)$  only for small values of  $p$ , that is  $\Delta p/p \sim r > 0$ . Thus if we want to let the population grow by 5% for a time interval only for small  $p$ , then we choose  $r$  to be 0.05. For some limiting value  $p = K > 0$  the population may not grow. For even greater values it may even decrease, for instance if the resources for the feeding of the population are limited, or if individuals in a large population are obstacles to each other etc.

Assume that the values  $y_n = \Delta p(n)/p(n)$  change linearly in  $p(n)$ . Graphically we can imagine this dependence as a line in the plane of the variables  $p$  and  $y$ . This line passes through the point  $[0, r]$ , so that  $y = r$  when  $p = 0$ . This line also passes through  $[K, 0]$  since this gives the second condition, namely that when  $p = K$  the population does not change. Thus we set

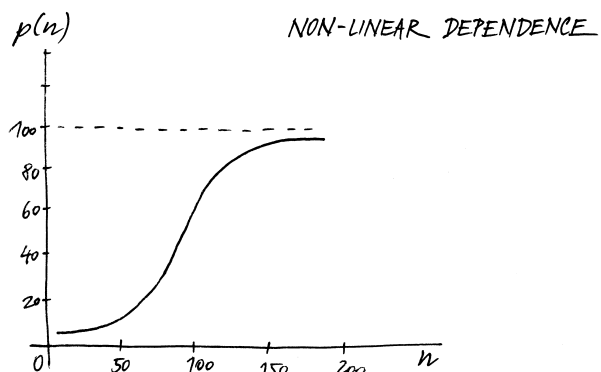
$$y = -\frac{r}{K}p + r.$$

By setting  $y = y_n = \Delta p(n)/p(n)$  and  $p = p(n)$  we obtain

$$\frac{p(n + 1) - p(n)}{p(n)} = -\frac{r}{K}p(n) + r.$$

By multiplying, we obtain a difference equation of first order with  $p(n)$  present as both a first and a second power.

$$(1) \quad p(n + 1) = p(n) \left( 1 - \frac{r}{K}p(n) + r \right).$$



```
f = y(n+1) - 3*y(n)/2 - 1
initial = {y(1):1, y(2):5/2}
```

To get the solution we finally type `rsolve(f, y(n), initial)`. This gives the answer  $2 * (3/2) * n - 2$ .<sup>5</sup> □

In Chapter 3 we will study recurrences in a more systematic way. Now we continue our observations with a very characteristic example, the Fibonacci numbers, which we will meet several times, see e.g., 3.B.1.

These numbers are defined by the recurrence relation using two preceding values:  $F_n = F_{n-1} + F_{n-2}$  ( $n \geq 3$ ) with the initial conditions  $F_1 = 1$  and  $F_2 = 1$  (or the same with  $F_0 = 0$ ). Hence the very first numbers in the Fibonacci sequence  $(F_n)_{n \in \mathbb{N}}$  are 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, etc.<sup>6</sup>

We shall call such equations (second order) linear difference equation with constant coefficients in Chapter 3, and we will learn how to solve them.

**1.B.4. Fibonacci numbers.** A 2-composition of a positive integer  $n$  is a representation of  $n$  as an ordered sum of numbers 1 and 2. Notice that the order indeed matters; For instance  $3 = 1 + 1 + 1$ ,  $3 = 1 + 2$  and  $3 = 2 + 1$ , hence the integer 3 has three distinct 2-compositions. Let  $C_n$  be the number of 2-compositions of  $n \in \mathbb{N}$ . Show that  $C_n = F_{n+1}$  for  $n \in \mathbb{N}$ . Having Sage as a tool, test (or prove) the claim, too.

**Solution.** Notice,  $C(0) = 1 = F(1)$  due to our conventions, since there is only the empty sum of interest providing the result. Thus, we may omit  $n = 0$  from our considerations. Let us denote by  $C_n(1)$  and  $C_n(2)$  the number of 2-compositions of  $n$ , ending with 1 and 2, respectively. For  $n = 1$  we have  $C_1(1) = 1$  and  $C_1(2) = 0$ , so  $C_1 = C_1(1) + C_1(2) = 1 + 0 = 1 = F_2$ . For  $n = 2$  we have  $C_2(1) = 1$  and  $C_2(2) = 1$ , so  $C_2 = C_2(1) + C_2(2) = 1 + 1 = 2 = F_3$ . Hence the claim holds for  $n = 1, 2$ . Suppose now that the integer  $n$  satisfies  $n \geq 3$ . Assume first that the 2-composition of  $n$  ends to 1. In this case, we can determine a 2-composition of the integer  $n - 1$  by deleting this 1. However, by adding a 1 in a 2-composition of  $n - 1$ , we will get a 2-composition of  $n$  that ends by 1. Thus we conclude that  $C_n(1) = C_{n-1}$ . Similarly, if the 2-composition of  $n$  ends by 2, then we get a 2-composition of the integer  $n - 2$  by deleting this 2. However, by adding a 2, or two ones, in the end of a 2-composition of  $n - 2$  we will result to a 2-composition of  $n$  ending either with 2 or with 1. Since the latter case has been counted already above, we conclude  $C_n(2) = C_{n-2}$ . Altogether we

<sup>5</sup>Recall that `a * n` in Sage means  $a^n$ .

<sup>6</sup>The Fibonacci numbers have a long and rich history, and they appear in many applications. These numbers were first discussed by Acharya Pingala, an Indian poet and mathematician (around 200 BC) when counting the possible patterns for poetry forms based on syllables of two lengths. The Italian mathematician Leonardo di Pisa (also called Bonacci, Pisano, or Fibonacci) explained them in his famous book Liber Abaci to the western community in 1202. Here these numbers were used to explain the rabbit reproduction question. They also arise in combinatorics and graph theory, and we shall meet them several times, e.g. in Chapters 3 and 13.

Try to think through the behaviour of this model for various values of  $r$  and  $K$ . In the diagram we can see the results for parameters  $r = 0.05$  (that is, five percent growth in the ideal state),  $K = 100$  (resources limit the population to the size 100), and as  $p(0) = 2$  we have initially two individuals.



Note that the original almost exponential growth slows down later. The population size approaches the desired limit of 100 individuals. For  $p$  close to one and  $K$  much greater than  $r$ , the right side of the equation (1) is approximately  $p(n)(1 + r)$ . That is, the behaviour is similar to that of the Malthusian model. On the other hand, if  $p$  is almost equal to  $K$ , the right hand side of the equation is approximately  $p(n)$ . For an initial value of  $p$  greater than  $K$  the population size will decrease. For an initial value of  $p$  less than  $K$  the population size will increase.<sup>1</sup>

### 3. Combinatorics

A typical “combinatorial” problem is to count in how many ways something can happen. For instance, in how many ways can we choose two different sandwiches from the daily offering in a grocery shop?



In this situation we need first to decide what we mean by different. Do we then allow the choice of two “identical” sandwiches? Many such questions occur in the context of card games and other games.

The solution of particular problems, usually involves either some multiplication of particular results (if the individual possibilities are independent) or some addition (if their appearance is disjoint). This is demonstrated in many examples in the problem column (cf. several problems starting with 1.D.2).

**1.3.1. Permutations.** Suppose we have a set of  $n$  (distinguishable) objects, and we wish to arrange them in some order. We can choose a first object in  $n$  ways, then a second in  $n - 1$  ways, a third in  $n - 2$  ways, and so on, until we choose the last object for which there is only one choice. The total number of possible arrangements is the product of these, hence there are exactly  $n! = n(n - 1)(n - 2) \dots 3 \cdot 2 \cdot 1$  distinct orders of the objects. Each ordering of the elements of a set  $S$  is called a *permutation* of the elements of  $S$ . The number of permutations on a set with  $n$  elements is  $n!$ .

We can identify the elements in  $S$  by numbering them (using the digits from one to  $n$ ), that is, we identify  $S$  with the set  $S = \{1, \dots, n\}$  of  $n$  natural numbers. Then the permutations correspond to the possible orderings of the numbers from one to  $n$ . Thus we have an example of a simple mathematical theorem and this discussion can be considered to be its proof.

<sup>1</sup>This model is called the *discrete logistic model*. Its continuous version was introduced already in 1845 by Pierre François Verhulst. Depending on the proportions of the parameters  $r$ ,  $K$  and  $p(0)$ , the behaviour can be very diverse, including chaotical dynamics. There is much literature on this model.

get  $C_n = C_n(1) + C_n(2) = C_{n-1} + C_{n-2}$ , and this coincides with the Fibonacci recurrence. Thus, the difference may be only in the initial conditions, but we have verified already  $C_1 = F_2$  and  $C_2 = F_3$ , and this concludes the proof.

To test  $C_n = F_{n+1}$  via Sage we will use the observed recurrence relation together with the initial conditions, which we present via Sage. Notice, how Sage imports tools from Python packages. Here Sage learns how to solve recurrences. Of course,  $c(n)$  would be listed very fast, after being resolved in the 6th line:

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n)-a(n-1)-a(n-2)
initial = {a(1):1, a(2):2}
c(n)=rsolve(f, a(n), initial)
k=var("k")
bool(c(k)==fibonacci(k+1)
    for k in (0 .. 120))
```

This returns True and so verifies our claim. Inside the cell we used the function `fibonacci(n)`, which returns the  $n$ th Fibonacci number, and the class `bool`, which is for comparing logical propositions. Later in 1.G.20 the reader can find a visual presentation of  $C_n$ .  $\square$

*Recurrence relations can be much more complicated than those of first or second order. To highlight the case, we present an example of order 3 and another one including more equations with more variables. In the latter case, although we are not able to evaluate the arbitrary term in the presented sequence  $P_{(k,l)}$  explicitly, we succeed in answering the question using the relevant recurrence relations. This is an example of the so called partial difference equations, since the terms of the sequence are indexed by two independent variables  $k, l$ .*

*Further applications of recurrence relations are in the final section of this chapter.*

**1.B.5.** How many words of length 12 can be constructed by using only the letters  $A$  and  $B$ , under the assumption that they do not contain a sentence of the form  $BBB$ ?

**Solution.** Let  $a_n$  denote the number of words of length  $n$  consisting of the letters  $A$  and  $B$ , but without  $BBB$  as a sentence in between. The words of length  $n > 3$  that satisfy the given condition, either end with one letter  $A$ , or with two letters  $AB$ , or with three letters  $ABB$  (notice that the  $AA$  two letters end has been covered, the  $ABB$  end is the only option for the  $BB$  end, while the remaining  $BAB$  three last letters are covered already). We enumerate  $a_{n-1}$  words ending with an  $A$  (observe that preceding the last  $A$ , it can be an arbitrary word of length  $n - 1$  satisfying our condition). Analogously for the two remaining groups. Thus  $a_n$  ( $n > 3$ ) satisfies the recurrence relation

$$a_n = a_{n-1} + a_{n-2} + a_{n-3}$$

NUMBER OF PERMUTATIONS

**Proposition.** The number  $p(n)$  of distinct orderings of a finite set with  $n$  elements, is given by the factorial function:

$$(1) \quad p(n) = n!$$

Suppose  $S$  is a set with  $n$  elements. Suppose we wish to choose and arrange in order just  $k$  of the members of  $S$ , where  $1 \leq k \leq n$ . This is called a  $k$ -permutation without repetition of the  $n$  elements. The same reasoning as above shows that this can be done in

$$v(n, k) = n(n-1)(n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

ways. The right side of this result also makes sense for  $k = 0$ , (there is just one way of choosing nothing), and for  $k = n$ , since  $0! = 1$ .

Now we modify the problem, this time where the order of selection is immaterial.

**1.3.2. Combinations.** Consider a set  $S$  with  $n$  elements. A  $k$ -combination of the elements of  $S$  is a selection of  $k$  elements of  $S$ ,  $0 \leq k \leq n$ , when order does not matter.



For  $k \geq 1$ , the number of possible results of a subsequential choosing of our  $k$  elements, is  $n(n-1)(n-2) \cdots (n-k+1)$  (a  $k$ -permutation). We obtain the same  $k$ -tuple in  $k!$  distinct orders. Hence the number of  $k$ -combinations is

$$\frac{n(n-1)(n-2) \cdots (n-k+1)}{k!} = \frac{n!}{(n-k)!k!}$$

If  $k = 0$ , the same formula is still true, since  $0! = 1$ , and there is just one way to select all  $n$  elements.

COMBINATIONS

**Proposition.** The number  $c(n, k) = \binom{n}{k}$  of combinations of  $k$ -th degree among  $n$  elements, where  $0 \leq k \leq n$ , is

$$(1) \quad c(n, k) = \frac{n(n-1) \cdots (n-k+1)}{k(k-1) \cdots 1} = \frac{n!}{(n-k)!k!}$$

We pronounce the *binomial coefficient*  $\binom{n}{k}$  as “ $n$  over  $k$ ” or “ $n$  choose  $k$ ”. The name stems from the *binomial expansion*, which is the expansion of  $(a+b)^n$ . If we expand  $(a+b)^n$ , the coefficient of  $a^k b^{n-k}$  is the number of ways to choose a  $k$ -tuple from  $n$  parentheses in the product (from these parentheses, we take  $a$ , from the others, we take  $b$ ). Therefore we have

$$(2) \quad (a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$$

Note that only distributivity, commutativity and associativity of multiplication and summation was necessary. The formula (2) therefore holds in every commutative ring.

with  $a_1 = 2$ ,  $a_2 = 4$ , and  $a_3 = 7$ . Using this relation we can now compute  $a_{12}$ . The answer is 1705. Of course, we can get this by simple modification of the previous Sage cell, give it a try!  $\square$

**1.B.6.** After the first quarter, the score of a basketball match between the national teams of Russia and the Czech Republic is 12 : 9 for the Czech team. In how many ways could the score have developed?

**Solution.** We can divide all possible evolutions of the quarter with the final score  $k : l$  into six mutually exclusive possibilities, according to which team scored, and how much was it worth (1, 2 or 3 points). If we denote by  $P_{(k,l)}$  the number of ways in which the score could have developed for a quarter that ended with  $k : l$ , then for  $k, l \geq 3$  the following recurrence relation holds:

$$P_{(k,l)} = P_{(k-3,l)} + P_{(k-2,l)} + P_{(k-1,l)} + P_{(k,l-1)} + P_{(k,l-2)} + P_{(k,l-3)}.$$

Using the symmetry of the problem,  $P_{(k,l)} = P_{(l,k)}$ . In addition, for  $k \geq 3$  we see that

$$\begin{aligned} P_{(k,2)} &= P_{(k-3,2)} + P_{(k-2,2)} + P_{(k-1,2)} + P_{(k,1)} + P_{(k,0)}, \\ P_{(k,1)} &= P_{(k-3,1)} + P_{(k-2,1)} + P_{(k-1,1)} + P_{(k,0)}, \\ P_{(k,0)} &= P_{(k-3,0)} + P_{(k-2,0)} + P_{(k-1,0)}. \end{aligned}$$

These relations, along with the initial condition, give  $P_{(0,0)} = 1$ ,  $P_{(1,0)} = 1$ ,  $P_{(2,0)} = 2$ ,  $P_{(3,0)} = 4$ ,  $P_{(1,1)} = 2$ , and

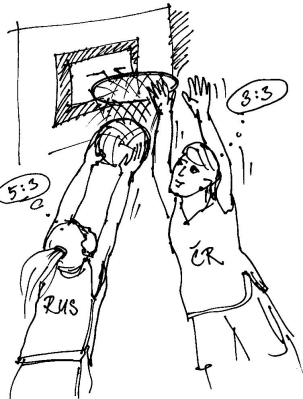
$$\begin{aligned} P_{(2,1)} &= P_{(1,1)} + P_{(0,1)} + P_{(2,0)} = 5, \\ P_{(2,2)} &= P_{(0,2)} + P_{(1,2)} + P_{(2,1)} + P_{(2,0)} = 14. \end{aligned}$$

Therefore, by repeatedly using the above equations, we arrive at nearly 500 million options, since  $P_{(12,9)} = 497178513$ .  $\square$

### C. Combinatorics

*In this section we use natural numbers to label items that may or may not be distinguishable and address questions about the number of ways certain events can occur. Let us begin with a straightforward problem involving permutations, which will demonstrate the usefulness of mathematical software packages, such as Matlab, Maple, Sage, etc.*

*We expect the reader to be familiar with the basic concepts as permutations, combinations, etc., as wells as adding or multiplying the numbers of independent options. If not, have a look into the other column in 1.3.1.*



We present a few simple propositions about binomial coefficients – another simple example of a mathematical proof. If needed, we define  $\binom{n}{k} = 0$  whenever  $k < 0$  or  $k > n$ .

**1.3.3. Proposition.** For all non negative integers  $n$ , we have

- (1)  $\binom{n}{k} = \binom{n}{n-k} \quad 0 \leq k \leq n.$
- (2)  $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1} \quad 0 \leq k \leq n - 1.$
- (3)  $\sum_{k=0}^n \binom{n}{k} = 2^n$
- (4)  $\sum_{k=0}^n k \binom{n}{k} = n2^{n-1}.$

**PROOF.** The first formula in the proposition is immediate directly from the formula 1.3.2(1). If we expand the right-hand side of (2), we obtain

$$\begin{aligned} \binom{n}{k} + \binom{n}{k+1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k-1)!} \\ &= \frac{(k+1)n! + (n-k)n!}{(k+1)!(n-k)!} \\ &= \frac{(n+1)!}{(k+1)!(n-k)!} \end{aligned}$$

which is the left-hand side of (2).

In order to prove (3), we use *mathematical induction* again. Mathematical induction consists of two steps. In the initial step, we establish the claim for  $n = 0$  (in general, for the smallest  $n$  the claim should hold for). In the inductive step we assume that the claim holds for some  $n$  (and all smaller numbers). We use this to prove that this implies the claim for  $n + 1$ . The principle of *mathematical induction* then asserts that the claim holds for every  $n$ .

The claim (3) clearly holds for  $n = 0$ , since  $\binom{0}{0} = 1 = 2^0$ . It holds also for  $n = 1$ . Now assume that the claim holds for some  $n \geq 1$ . We must prove the corresponding claim for  $n + 1$  using the claims (2) and (3). We calculate

$$\begin{aligned} \sum_{k=0}^{n+1} \binom{n+1}{k} &= \sum_{k=0}^{n+1} \left( \binom{n}{k-1} + \binom{n}{k} \right) \\ &= \sum_{k=-1}^n \binom{n}{k} + \sum_{k=0}^{n+1} \binom{n}{k} = 2^n + 2^n = 2^{n+1}. \end{aligned}$$

Note that the formula (3) gives the number of all subsets of an  $n$ -element set, since  $\binom{n}{k}$  is the number of all subsets of size  $k$ . Note also that (3) follows from 1.3.2(2) by choosing  $a = b = 1$ .

To prove (4) we again employ induction, as we did in (3). For  $n = 0$  the claim clearly holds. The inductive assumption says that (4) holds for some  $n$ . We calculate the corresponding sum for  $n + 1$  using (2) and the inductive assumption. We obtain



the number of permutations of seven elements. By symmetry, the number of all orderings where A speaks directly after B is also 7!. Since the number of all possible orderings of eight speakers is 8!, the solution is  $8! - 2 \cdot 7!$ .  $\square$

**1.C.3.** How many rearrangements of the letters of the word “problem” are there, such that:

- (a) The letters b, r are next to each other;
- (b) The letters b, r are not next to each other.  $\circ$

**1.C.4.  $k$ -permutations and  $k$ -combinations.** Determine the number of different permutations that admits a code with 5 digits in a specific order, under the assumption that every digit can be used only once. Deduce that 5-digit codes of such type are much more secure than the corresponding 4-digit codes. Finally, what is the situation if the order of digits in the code does not matter?

**Solution.** We are interested in codes with 5 digits, where the order of the digits is crucial and every digit appears only once. In other words, we have the set of digits, say  $S = \{0, 1, \dots, 9\}$ , and we want to choose and arrange in order 5 elements of  $S$ . In terms of the paragraph 1.3.4, this is a 5-permutation of 10 elements without repetition, which can be done in

$$u(10, 5) = \frac{10!}{(10-5)!} = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 = 30240$$

ways. In Sage we can compute  $u(10, 5)$  by typing

```
factorial(10)/factorial(5)
```

For a 4-digit code with the same characteristics, the corresponding number of permutations is given  $u(10, 6) = 5040$ , hence the claim. (Of course, we see from the formula that the letter number must be 6 times smaller than the first one!)

Assume now that the order does not play some role. Then, codes of the form 12345 and 54123, for instance, represent the same code, hence finally one will get much less possible permutations (codes of this type are very unsafe). In particular, this case yields the definition of a 5-combination of 10 elements, whose number according to 1.3.2 is given by

$$c(10, 5) = \binom{10}{5} = \frac{10!}{5!(10-5)!} = \frac{30240}{120} = 252.$$

Sage provides the command `binomial(n,k)` for the  $k$ -combination of  $n$  elements which in our case applies as

```
binomial(10, 5)
```

In such terms, the very first result about  $u(10, 5)$  appears also by the command `binomial(10,5) * factorial(5)`.  $\square$

**1.C.5. Remark.** Notice that Sage does not have a direct command for  $k$ -permutations. Let us construct one, by introducing a function called `uperm(n, k)`:<sup>7</sup>

<sup>7</sup>Note that usually the cells that we provide in Sage can be copy-pasted without extra editing being necessary in the editor of Sage. However, when we introduce functions as below, the reader should be careful and type the command in the editor exactly as it appears here, since the spaces play some role.

$k$ -PERMUTATIONS WITH REPETITIONS

$$V(n, k) = n^k.$$

If we are interested in a choice of  $k$  elements without taking care of order, we speak of  $k$ -combinations with repetitions. At first sight, it does not seem to be easy to determine the number. We reduce the problem to another problem we have already solved, namely combinations without repetitions:

COMBINATIONS WITH REPETITIONS

**Theorem.** The number of  $k$ -combinations with repetitions from  $n$  elements equals for every  $k \geq 0$  and  $n \geq 1$

$$\binom{n+k-1}{k}.$$

**PROOF.** Label the  $n$  elements as  $a_1, a_2, \dots, a_n$ . Suppose each element labeled  $a_i$  is selected  $k_i$  times,  $0 \leq k_i \leq k$ , so that  $k_1 + k_2 + \dots + k_n = k$ . Each such selection can be paired with the sequence of symbols  $*$  and  $|$  where each  $*$  represents one selection of an element and individual boxes are separated by  $|$  (therefore there are  $n-1$  of them).

The number of  $*$  in the  $i$ th box is equal to  $k_i$ , so we obtain the sequence

$$\underbrace{* \dots *}_{k_1} | \underbrace{* \dots *}_{k_2} | \dots | \underbrace{* \dots *}_{k_n}.$$

The other way around, from any such sequence we can determine the number of selections of any element (e.g. the number of  $*$  before first  $|$  determines  $k_1$ ).

Having altogether  $k$  symbols  $*$  and  $n-1$  separators  $|$  we see that there are

$$\binom{n+k-1}{n-1} = \binom{n+k-1}{k}$$

possible sequences and therefore also the same number of the required selections.  $\square$

4. Probability

Now we are going to discuss the last type of function description, as listed in the very end of the subsection 1.1.5. Thus, instead of assigning explicit values of a function, we shall try to describe the probabilities of the individual options.



**1.4.1. What is probability?** As a simple example we can use common six-sided dice throwing, with sides labelled as

- 1, 2, 3, 4, 5, 6.

If we describe the mathematical model of such throwing with a “fair” dice, we expect by symmetry that every side occurs with the same frequency. We say that “every side occurs with the probability  $1/6$ ”.

```
def uperm (n, k) :
    a = factorial (k) * binomial (n, k)
    return (a)
```

In such terms we can directly compute  $u(n, k)$  for any  $n, k$ , with  $k = 0, 1, 2, \dots$ , by typing in the editor of Sage the command `uperm(n, k)`. For instance, the command `uperm(10, 5)` returns 30240, `uperm(10, 1)` returns 10, `uperm(10, 2)` returns 90, `uperm(10, 11)` returns 0 (as it should be since  $k > n$  in this case), etc.

**1.C.6.** Determine the number of 4-digit codes, composed of the digits 1, 3, 5, 6, 7 and 9, under the condition that no digit occurs more than once.

**Solution.** We have 6 distinct letters at our disposal, and we ask: How many distinct (ordered) 4-tuples can be chosen from them? Obviously the result is

$$u(6, 4) = \frac{6!}{(6-4)!} = 6 \cdot 5 \cdot 4 \cdot 3 = 360.$$

The reader may verify the computation in Sage with the function `uperm(n, k)` constructed above, that is, via the command `uperm(6, 4)`.  $\square$

**1.C.7.** Six men in a meeting shake mutually their hands. How many handshakes will happen in total?

**Solution.** We understand that each couple of men shake their hands. Thus, the number of handshakes equals the number of ways of choosing an unordered pair among 6 elements, this is the combinations  $c(6, 2)$ . Let us present the answer via Sage:

```
c62=binomial(6, 2); c62
```

which gives 15.  $\square$

*To summarize,  $k$ -permutations  $u(n, k)$  take care of the order while combinations  $c(n, k)$  do not. Moreover,  $u(n, k) = c(n, k)k!$ . We have noticed that permutations and combinations are most useful when counting possibilities. Let us finalize this paragraph by treating examples where repetitions occur. Notice how the principle of “inclusion and exclusion” may be useful.*

**1.C.8.  $k$ -permutations with repetition.** The Greek alphabet consists of 24 letters. How many words of exactly five letters can be composed in it? (Disregarding whether the words have some actual meaning or not).

**Solution.** For each of the five positions in the word we have 24 possibilities, since the letters can repeat. According to the discussion in 1.3.4, this is nothing but a 5-permutation with repetitions. Hence the total number of words that can be composed is given by  $V(24, 5) = 24^5 = 7962624$ . In Sage we compute this by typing `24 ^ 5` or `24 * * 5`.  $\square$

But throwing some less symmetric version of a dice with six faces, the actual probabilities of the individual results might be quite different. Let us build a simple mathematical model for this. We shall work with the parameters  $p_i$  for the probabilities of individual sides with two requirements. These probabilities have to be non-negative real numbers and their sum is one, i.e.

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1.$$

At this time, we are not concerned about the particular choice of the specific values  $p_i$ , they are given to us. Later on, in chapter 10, we shall link probability with mathematical statistics and then we shall introduce methods how to discuss reliability of such a model for a specific real dice.

**1.4.2. Classical probability.** Let us come back to the mathematical model for the fair dice. We consider the sample space  $\Omega = \{1, 2, 3, 4, 5, 6\}$  of all possible elementary events (each of them corresponding to one possible result of the experiment of throwing the dice). Then we can consider any event as a given subset  $A$  of  $\Omega$ . For example  $A = \{1, 3, 5\}$  describes the result of getting odd number on the resulting side (we count the labels on the sides of the dice). Similarly, the set  $B = A^c = \{2, 4, 6\} = \Omega \setminus A$  is the complementary event of getting even numbered points. The probability of both  $A$  and  $B$  will be  $1/2$ . Indeed,  $|A|/|\Omega| = 1/2$ , where  $|A|$  means the number of elements of a set  $A$ .

This leads to the following obvious generalization:

#### CLASSICAL PROBABILITY

Let  $\Omega$  be a finite set with  $n = |\Omega|$  elements. The *classical probability* of the event corresponding to any subset  $A \subset \Omega$  is defined as

$$P(A) = \frac{|A|}{|\Omega|}.$$

Such a definition immediately allows us to solve problems related to throwing several fair dice simultaneously. Indeed, we may treat this as throwing independently one dice many times and thus multiplying the probabilities. For example, the event of getting an odd sum of points on two dice is given by adding the probabilities of having an even number on the first one and odd number on the second one and vice versa. Thus the probability will be twice  $1/2 \cdot 1/2$ , which is  $1/2$  as expected.

**1.4.3. Probability space.** Next, we formulate a more general concept of probability covering also the unfair dice example above.



We shall need a finite set  $\Omega$  of all possible states of a system (e.g. results of an experiment), which we call the *sample space*.

Further, the space of all possible *events* is given as the set  $\mathcal{A}$  of all subsets in  $\Omega$ . Finally, we need the function describing the probabilities of occurrence of individual events:



**1.C.9.** In how many ways can we insert five golf balls into five holes, with one ball into every hole, assuming that we have four identical white balls, four identical blue balls and three identical red balls?

**Solution.** First it is convenient to solve the problem for the case where we have five balls of every colour. This amounts to a free choice of five elements from three possibilities (there is a choice out of three colours for every hole), that is, permutations with repetitions. So the total ways are given by  $V(3, 5) = 3^5 = 243$ . Next, to answer our task it is sufficient to subtract the configurations which are excluded:

- all chosen balls are of one colour; there are 3 such cases,
  - exactly four red balls were chosen; there are  $2 \cdot 5 = 10$ .
- Indeed, in the latter event, one first fixes the colour of the non-red ball (two options), and then the hole where this ball is in (five independent options). Hence, in total there are  $3^5 - 3 - 10 = 230$  possible ways.  $\square$

**1.C.10.** Determine the number of ways of placing the white tower and black tower on the chessboards (of size  $8 \times 8$ ), that are neither in the same column nor in the same row.  $\circ$

**1.C.11.  $k$ -combinations with repetition.** In how many ways can we equip three distinct envelopes with five identical stamps and five identical letters. (Yes, we may glue many or no stamps onto the same envelop, etc.) Next provide an answer for the same task, under the condition that no envelope stays empty.

**Solution.** First compute the number of choices ignoring the non-emptiness condition. It is an example of 5-combinations with repetition from 3 elements, and since we glue the stamps and insert the letters independently, we have got  $C(3, 5) \cdot C(3, 5) = c(7, 2)^2 = 441$  different ways.

Next we have to subtract the choices with exactly one envelope empty, and finally with two empty ones (we still may distribute the stamps freely). We thus obtain (in the second term we choose the empty envelop and distribute the letters, and we have to correct the multiple appearances)  $C(3, 5) - 3(C(2, 5) + 3) - 3C(3, 5) = \binom{7}{2} - 3\binom{6}{1} + 3 = 6$ . Thus the resulting number is  $6C(3, 5) = 126$ .

If our interpretation was that the stamps are glued only on the non-empty envelopes, the result would be the square of the latter number of choices for envelopes, i.e., 36.  $\square$

### D. Probability

*We proceed with simple exercises related to the concept of classical probability. In many cases we are dealing with experiments having only a finite number of outcomes and we are interested in whether or not the outcome belongs to a subset of favourable outcomes. Then, the probability we are trying to determine equals the number of favourable outcomes divided by the total number of all possible outcomes. Classical probability can be used when we assume, or know, that each*



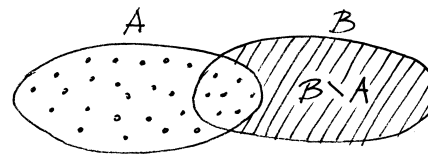
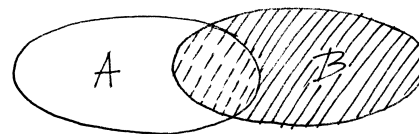
### PROBABILITY FUNCTION

Let us consider a non-empty fixed sample space  $\Omega$ . The probability function  $P : \mathcal{A} \rightarrow \mathbb{R}$  satisfies

- (1)  $P(\Omega) = 1$
- (2)  $0 \leq P(A)$  for all events  $A$
- (3)  $P(A \cup B) = P(A) + P(B)$  whenever  $A \cap B = \emptyset$ .

Notice that the intersection  $A \cap B$  describes the simultaneous appearance of both events, while the union  $A \cup B$  means that at least one of events  $A$  and  $B$  appear. The event  $A^c = \Omega \setminus A$  is called the *complementary event*.

### PROBABILITY



$$P(A \cup B) = P(A) + P(B \setminus A)$$

There are some further straightforward consequences of the definition for all events  $A, B$ :

- (4)  $P(A) = 1 - P(A^c)$
- (5)  $P(\emptyset) = 0$
- (6)  $P(A) \leq 1$  for all events  $A$
- (7)  $P(A) \leq P(B)$  whenever  $A \subset B$
- (8)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The proofs are all elementary. For example,  $A \cup (A^c) = \Omega$  and thus (3) implies (4).

Similarly, we can write  $A = (A \setminus B) \cup (A \cap B)$  and  $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$  with disjoint unions of sets on the right hand sides. Thus,  $P(A) = P(A \setminus B) + P(A \cap B)$  and  $P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B)$  by (3), which implies the last equality. The remaining three claims are even simpler.

All these properties correspond exactly to our intuition how probability should behave. Probability should be always a real number between zero and one. The event  $\Omega$  includes all possible results of the experiment, so it must have probability one. No result appears with probability zero, the probabilities of disjoint events should add, etc.

Of course, the classical probability on the sample space  $\Omega$  is an example of a probability function. In our more general model, the set of all subsets  $\mathcal{A}$  is closed upon union, intersection and taking the complement, and this has been essential

outcome has the same probability of happening (for instance, fair dice throwing, etc).

**1.D.1.** Prove that the following two events can happen with the same probability (independently of each other):

- (a) The roll of a dice results in a number greater than 4;
- (b) Throwing two dice, and assuming that their total sum equals to 7, at least one resulted in four.

**Solution.** For the first case, there are six possible outcomes (the set  $\{1, 2, 3, 4, 5, 6\}$ ). Two are favourable ( $\{5, 6\}$ ). Thus the probability is  $2/6 = 1/3$ . For the second event, one can apply classical probability (where the condition is interpreted as restriction of the probability space). The space has 6 elements, and exactly 2 of those are favourable to the given event. Thus, the answer is again  $2/6 = 1/3$ .

Notice, that if we treat rolling of two dices as two independent rolling of one, and view the results independent of the order (i.e., just the total), then the probability of the sums is not equal. For instance 12 and 2 have got probability  $1/36$  while 10 appears with probability  $3/36$ , etc. Thus, throwing two dices is not a classical probability case in general, but we were right with treating the resulting sum 7 as classical probability anyhow. Why?  $\square$

**1.D.2.** Helen wants to give to John and Mary five pears and six apples. If we consider the pears to be indistinguishable and the same for the apples, what is the probability that either John or Mary get nothing?

**Solution.** Let us assume that the pears and apples are distributed independently and each option has the same probability. Of course, one of the two mentioned persons could then get nothing. We have to count in how many ways the distributions happen.

The five pears can be divided in six ways (it is determined by the number of pears given to John, the rest goes to Mary.) Similarly, the six apples can be divided in seven ways. These divisions are independent, hence we can apply the product rule, which states that the frequency of two independent events occurring together can be calculated by multiplying the individual values, i.e., we have got  $6 \cdot 7 = 42$  options in total. Two of them result in John or Mary getting nothing. Thus the probability is  $1/21$  (relatively high).

Notice, how much different this is from dealing with 11 distinguishable objects, where 11 independent tosses of the same coin with always the same result would result in probability  $2 \cdot 2^{-11}$  (extremely low).  $\square$

**1.D.3.** We choose randomly a group of five people from a group of eight men and four women. What is the probability that there are at least three women in the chosen group?  $\circ$

**1.D.4. Remark.** The classical probability concept can be often applied, but we have to be careful. Imagine we should compute the probability that the reader of this remark will win at least 25 million euro in EuroLotto during the next week. First, notice the formulation of the problem is incomplete. For

in our exposition above. This will continue in all our discussion on probability in the sequel, where we shall have to allow for more general spaces of events  $\mathcal{A}$  in the sets of all subsets in the sample space. We will return to this and more serious generalizations in chapter 10.

**1.4.4. Summing probabilities.** By using mathematical induction, the additivity of probability is easily extended to any (finite) number of mutually exclusive events  $A_i \subset \Omega$ ,  $i = 1, \dots, n$ . That is,

$$P(\cup_{i \in I} A_i) = \sum_{i=1}^n P(A_i),$$

whenever  $A_i \cap A_j = \emptyset$ , for all  $i \neq j$ ,  $i, j = 1, \dots, n$ . Indeed, 1.4.3(3) is the result for  $n = 2$ . If we assume the validity of the formula for some fixed  $n$ , then the union of any  $n + 1$  events  $A_0, A_1, \dots, A_n$  can be split into the union of  $A_0$  and  $A_1 \cup \dots \cup A_n$ . Then by the induction assumption, together with 1.4.3(3) again, the result follows.

In general, the summing of probabilities of event occurrences is much more difficult. The problem is that whenever the events are mutually compatible, the possible results in their intersection are counted multiple times.

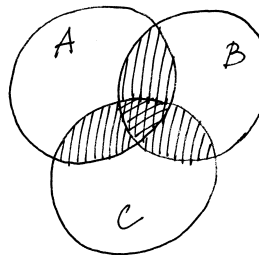
We have seen the simplest case of two mutually compatible events  $A$  and  $B$  in 1.4.3(8). For classical probability, it reduces just to counting elements in subsets. Indeed, those elements that belong to both the sets  $A$  and  $B$  count in the formula  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$  twice and thus we have to subtract them once.

Now, we look at the general case. The approach of interactive inclusion and exclusion (potentially too many) elements in some count is a standard method in combinatorics known as the *inclusion-exclusion principle*. We shall exploit this method in our general finite probability spaces.

As we shall see, this is an example of a mathematical theorem, where the hard part is to understand (and find) the formulation of the result. The proof is then relatively simple.



INCLUSION-EXCLUSION PRINCIPLE



The diagram explains the situation for three sets  $A, B, C$  for classical probability:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Clearly, the probabilities are given by first counting the elements in each set and adding. Then we subtract the sum

example, we could impose that the sample space is just “win” and “not win”, with the same probability 1/2. In reality, the basic condition for the classical probability is not valid (every elementary event must have the same probability), but even the elementary event has not been defined sufficiently yet.

Instead, suppose that EuroLotto has a daily draw with a jackpot of €25 000 000 for choosing 5 correct numbers in  $1, \dots, 50$ . Let us simplify our problem and assume we want to win the €25 000 000 jackpot at least once. Thus we want to see the probability that the elementary event, a single lotto card with 5 numbers winning a jackpot happens. We assume that all quintuples are drawn with the same probability.

Assuming for instance that the reader submits  $k$  lotto cards every day of the week, the probability of winning at least one jackpot during the week is  $\frac{7k}{\binom{50}{5}} = \frac{7k}{2118760}$ . For instance, for  $k = 7$  via Sage and the command

```
N(49/2118760)
```

we get the very small number 0.0000231267345050879.

**1.D.5.** There are  $2n$  seats in a row in a cinema. We randomly seat  $n$  men and  $n$  women in the row. What is the probability that no two persons of the same sex sit next to each other?

**Solution.** There are  $(2n)!$  possible seatings. We assume, each of them comes with the same probability. The number of seatings satisfying the given condition is  $2(n!)^2$ . For we have two ways for choosing the positions for men and thus also for women: either all men sit on odd-numbered places, and thus the women sit on even-numbered places, or vice versa. Among these places, both men and women are seated arbitrarily. Therefore, the resulting probability is

$$P(n) = \frac{2(n!)^2}{(2n)!}.$$

In particular,  $P(2) = 1/3$ ,  $P(5) = 1/126$  and  $P(8) = 1/6435$ . To verify this computation by Sage we use the cell:

```
n=var("n")
P(n) = 2*factorial(n)**2/factorial(2*n)
```

Typing  $P(2)$ ,  $P(5)$ , etc, we obtain the given values. □

**1.D.6.** Five persons enter an elevator in a building with eight floors. Each of them leaves the elevator at any floor with the same probability. What is then the probability, that

- all of them leave at sixth floor;
- all of them leave at the same floor;
- each of them leaves at a different floor.

**Solution.** The sample space of possible events (see 1.4.3), is the space of all possible ways of leaving the elevator by 5 people. There are  $8^5$  of them. Hence, in the first case there is only one favourable outcome, thus the probability is  $\frac{1}{8^5}$ . In the second case there are eight favourable outcomes, thus the probability is  $8 \frac{1}{8^5} = \frac{1}{8^4}$ . In the third case, the number of favourable outcomes is given by a five-element variation of eight elements (we choose five floors among eight where

of those in intersections of pairs of sets, since those elements are counted twice. But we must then add in the number of elements in the intersection of all three.

We shall now follow the same idea in order to write down the formula in the following theorem. It seems plausible that such a formula should work with proper coefficients of the sums of probabilities of intersections of more and more events among  $A_1, \dots, A_k$ , at least in the case of classical probability. The reader will perhaps appreciate that a quite straightforward mathematical induction will verify the theorem in full generality.

**1.4.5. Theorem.** Let  $A_1, \dots, A_k \in \mathcal{A}$  be arbitrary events over the sample space  $\Omega$  with a set of events  $\mathcal{A}$ . Then

$$\begin{aligned} P(\cup_{i=1}^k A_i) &= \sum_{i=1}^k P(A_i) - \sum_{i=1}^{k-1} \sum_{j=i+1}^k P(A_i \cap A_j) \\ &\quad + \sum_{i=1}^{k-2} \sum_{j=i+1}^{k-1} \sum_{\ell=j+1}^k P(A_i \cap A_j \cap A_\ell) \\ &\quad - \dots \\ &\quad + (-1)^{k-1} P(A_1 \cap A_2 \cap \dots \cap A_k). \end{aligned}$$

**PROOF.** For  $k = 1$  the claim is obvious. The case  $k = 2$  is the same as the equality 1.4.3(8), which we have already proved.

Assume that the theorem holds for any number of events up to  $k$ , where  $k \geq 1$ . Now we can work in the induction step with the formula for  $k + 1$  events, where the union of the first  $k$  of them are considered to be the  $A$  in the equation 1.4.3(8) and the remaining event is considered to be the  $B$ :

$$\begin{aligned} P(\cup_{i=1}^{k+1} A_i) &= P((\cup_{i=1}^k A_i) \cup A_{k+1}) \\ &= \sum_{j=1}^k \left( (-1)^{j+1} \sum_{1 \leq i_1 < \dots < i_j \leq k} P(A_{i_1} \cap \dots \cap A_{i_j}) \right) \\ &\quad + P(A_{k+1}) - P((A_1 \cup \dots \cup A_k) \cap A_{k+1}). \end{aligned}$$

This already resembles the formula for  $k + 1$  summed events. But in the first term, expressions containing  $A_{k+1}$  are missing. Also absent is a term allowing for the probability that all the events happen. On the other hand, the last expression should not be there. We can replace it by the expression

$$-P((A_1 \cap A_{k+1}) \cup \dots \cup (A_k \cap A_{k+1}))$$

and for this we can again use the induction, that is, the formula in the statement of the theorem. With a little patience (and a piece of paper long enough to write down all the expressions) we can check that this adds all the missing pieces. □

**1.4.6. Inclusion-exclusion principle.** As we have mentioned already, a special case of the previous theorem is the one of classical probability. There the probability of an event  $A$  is strictly proportional to the number of elements in  $A$  (which is just divided by the total size  $n$  of the sample space).



some person leaves the elevator and then we choose the order in which they leave the chosen floors). The probability is then

$$\frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4}{8^5} \doteq 0.20508.$$

In Sage we can obtain this result by using the function `uperm` introduced above. We present the corresponding cell:

```
def uperm(n, k):
    a = factorial(n)/factorial(n-k)
    return a
p=uperm(8,5)/8**5
print("Probability of each of them leaving
at a different floor:", round(p,5))
```

It returns (the round function rounds a number to a given precision in decimal digit):

```
Probability of each of them leaving
at a different floor: 0.20508
```

□

**1.D.7.** In a particular state there is a parliament with 200 members. Two major political parties in this state flip a coin during an “election” for every seat in the parliament. Each of the parties has associated one side of the coin. What is the probability that each of the parties gains 100 seats, assuming that the coin is “fair”?

**Solution.** There are  $2^{200}$  of possible results of the elections (considered to be a sequence of 200 results of flips). If each party is to obtain 100 seats, then there are exactly 100 tails and 100 heads in the sequence. There are  $\binom{200}{100}$  such sequences (since the sequence is uniquely determined by choosing 100 members of 200 possible, which will result in, say, tails). The resulting probability is

$$\frac{\binom{200}{100}}{2^{200}} = \frac{200!}{100! \cdot 100!} \doteq 0.0563.$$

In Sage we can do this computation by using the `binomial` and `round` functions.

```
ans = binomial(200, 100) / (2**200)
print(round(ans, 4))
```

□

**1.D.8.** In a box there are 9 red and 7 white balls. Sequentially we draw three balls (without returning). Determine the probability that the first two are red and the third is white. ○

*In many problems we use the “rule of complement”, computing the size of the complements of the desired subsets. This principle generalizes to the notion of “inclusion-exclusion principle”. Iterating, we add too much, push out unwanted counts, etc., see 1.4.6.*



Thus, in the formula from the previous theorem, all the probabilities give the sizes of the subsets involved, up to a common factor  $\frac{1}{n}$ .

In this way we can extract from the theorem 1.4.5 the following claim for the size of a general finite set  $M$  and its subsets  $A_1, \dots, A_k$ . As usual we let  $|M|$  denote the number of elements of the set  $M$ .

Of course for every finite set  $M$  and its subsets,

$$|M \setminus (\cup_{i=1}^k A_i)| = |M| - |\cup_{i=1}^k A_i|.$$

Now we use the previous theorem, and express the size of the union on the right side, and we obtain the theorem that is usually called

PRINCIPLE OF INCLUSION-EXCLUSION

$$|M \setminus (\cup_{i=1}^k A_i)| = |M| + \sum_{j=1}^k (-1)^j \sum_{1 \leq i_1 < \dots < i_j \leq k} |A_{i_1} \cap \dots \cap A_{i_j}|$$

The meaning of this result for the special case  $n = 3$  can be visualized easily, see the diagram in 1.4.5.

**1.4.7. Independent events.** Next, we wish to express possible dependencies among events in a given sample space  $\Omega$  with the probability function  $P$ . We say that the events  $A$  and  $B$  are *stochastically independent* if

$$P(A \cap B) = P(A) \cdot P(B).$$

This definition may remind us of our experiences in combinatorics when counting possibilities for independent choices. For example, dealing with a fair dice, we can define events  $A$  “odd number occurs”,  $B$  “the result is at least 3” and  $C$  “the result is at most 3”. The probabilities are  $P(A) = \frac{1}{2}$ ,  $P(B) = \frac{2}{3}$ ,  $P(C) = \frac{1}{2}$ ,  $P(A \cap B \cap C) = \frac{1}{6} = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{2} = P(A) \cdot P(B) \cdot P(C)$ , and taking pairs we have  $P(A \cap C) = \frac{1}{3} \neq \frac{1}{2} \cdot \frac{1}{2}$ ,  $P(A \cap B) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3}$ ,  $P(B \cap C) = \frac{1}{6} \neq \frac{2}{3} \cdot \frac{1}{2}$ . Notice, that the stochastic dependence of the pairs  $A, C$  and  $B, C$  corresponds well to our intuition (e.g. there are more odd numbers between the values 1, 2, 3 than between the numbers 4, 5, 6).

This example also shows that we have to be careful with more events. In general, mutually independent sets are defined in this way:

**Definition.** Consider an arbitrary probability function  $P$  on sample space  $\Omega$  and  $k$  events  $A_1, \dots, A_k$  in that space. We say that these events are *stochastically independent* (with respect to the probability function  $P$ ), if for any chosen events  $A_{i_1}, \dots, A_{i_\ell}$ ,  $1 \leq \ell \leq k$  we have



$$P(A_{i_1} \cap \dots \cap A_{i_\ell}) = P(A_{i_1}) \cdot \dots \cdot P(A_{i_\ell}).$$

Every subset of a set of stochastically independent events is also stochastically independent. Further, for any two

**1.D.9. Inclusion-exclusion principle.** A secretary has to send six letters to six different people. She puts the letters in the envelopes randomly. What is the probability that at least one person receives the correct intended letter?

**Solution.** We compute the probability of the complementary event, which is “no person receives the correct letter”. The sample space corresponds to all possible orderings of six envelopes. Let us denote both the letters and the envelopes by numbers from one to six. Then, all the favourable events (no letter is assigned to the corresponding envelope) correspond to such orderings of six elements, where the  $i$ th element is not at the  $i$ th place, for  $i = 1, \dots, 6$ . These are the permutations without a fixed point.

One can compute the number of such orderings using the so-called *inclusion-exclusion principle*, see Theorem 1.4.5 and its corollary 1.4.6. Here we will use the same notation. Let us denote by  $M_i$  the set of permutations, such that  $i$  is a fixed point (note that permutations in  $M_i$  can also have other fixed points). In such terms, the resulting number  $d$  of permutations without a fixed point is



$$d = 6! - |M_1 \cup \dots \cup M_6|.$$

On the other hand, the number of elements in the intersection  $M_{i_1} \cap \dots \cap M_{i_k}$ ,  $k = 1, \dots, 6$ , equals to  $(6-k)!$  (note that the order of the elements  $i_1, \dots, i_k$  is fixed, while the remaining  $6-k$  can be ordered arbitrarily). The inclusion-exclusion principle states that

$$|M_1 \cup \dots \cup M_6| = \sum_{k=1}^6 (-1)^{k+1} \binom{6}{k} (6-k)!.$$

Hence, we can compute the number  $d$ :

$$\begin{aligned} d &= 6! - \sum_{k=1}^6 (-1)^{k+1} \binom{6}{k} (6-k)! \\ &= \sum_{k=0}^6 (-1)^k \binom{6}{k} (6-k)! = 6! \sum_{k=0}^6 \frac{(-1)^k}{k!}. \end{aligned}$$

We deduce that the probability that no person receives “his” letter is  $\sum_{k=0}^6 \frac{(-1)^k}{k!} = \frac{53}{144}$ , and thus the probability we were asked for equals to  $1 - \frac{53}{144} = \frac{91}{144}$ . More than one half!  $\square$

**1.D.10.** Approximately 1200 persons die per year at the roads of the Czech Republic. Determine the probability that some person of a chosen group of 500 people will die in the following ten years in a traffic accident (assume that the total population is 10 million and every person has the same probability of dying in a traffic accident during one year).

**Solution.** Let us first count the probability that one randomly chosen person will not die in ten years in a traffic accident. The probability that he/she will not die in one year is  $(1 - \frac{12}{10^5})$ . Thus, the probability that he/she will not die in ten years is  $(1 - \frac{12}{10^5})^{10}$ . We deduce that the probability that in ten years

stochastically independent events we compute

$$\begin{aligned} P(A \cap B^c) &= P(A \setminus B) = P(A) - P(A \cap B) = \\ &= P(A)(1 - P(B)) = P(A)P(B^c). \end{aligned}$$

From there we can show that by exchanging one or more events in a set of stochastically independent events by their complements, we again obtain a set of stochastically independent sets.

Sometimes we need to compute the probability that at least one of the stochastically independent set of events occurs. That is, we want to compute  $P(A_1 \cup \dots \cup A_k)$ . In such a situation we can use the De Morgan laws for sets (check them yourselves!)

$$(\cup_{i \in I} A_i)^c = \cap_{i \in I} A_i^c, \quad (\cap_{i \in I} A_i)^c = \cup_{i \in I} A_i^c.$$

We obtain

$$\begin{aligned} P(A_1 \cup \dots \cup A_k) &= 1 - P(A_1^c \cap \dots \cap A_k^c) \\ &= 1 - (1 - P(A_1)) \cdot \dots \cdot (1 - P(A_k)) \\ &= 1 - \prod_{j=1}^k (1 - P(A_j)). \end{aligned}$$

**1.4.8. Conditional probability.** Often we want to restrict our attention only to events, which lie in a subspace  $H \subset \Omega$ . This means that the events in question will be the intersections  $A \cap H$  of the original events  $A$  with the subset  $H$ . Thus our new probabilities should be proportional to  $P(A \cap H)$ . We would like to have  $H$  in the role of the new sample space.



As an example, we might look again at the model of a fair dice and ask the question “what is the probability that by throwing two dice the result is twice 5, if we know that the sum of the results is 10?”. Of course, we are now having only the possibilities  $4 + 6$  (two times) and  $5 + 5$  (once). So the probability should be  $\frac{1}{3}$ , much greater then the probability  $\frac{1}{36}$  of the same event without any further condition.

Similar situations are reflected in the following definition:

#### CONDITIONAL PROBABILITY

**Definition.** Let  $H$  be an event with non-zero probability in the sample space  $\Omega$  with the probability function  $P$ . The *conditional probability*  $P(A|H)$  of the event  $A$  given  $H$  is defined by the formula

$$P(A|H) = \frac{P(A \cap H)}{P(H)}.$$

The event  $H$  is sometimes called the *hypothesis*.

As it is obvious from the definition, the hypothesis  $H$  with non-zero probability and the event  $A$  are (stochastically) independent if and only if  $P(A) = P(A|H)$ . The definition also directly implies the “theorem for product of probabilities”

none of the given 500 people will not die, occurs by the product rule. This is because the events are independent, hence we obtain  $(1 - \frac{12}{10^5})^{5000}$ . It follows that the probability of the complementary event, that is, some of the chosen people will die, equals to  $1 - (1 - \frac{12}{10^5})^{5000} \doteq 0.4512$ .  $\square$

**Remark.** The model we have used in the previous exercise to describe the given situation is only approximate. The complication arises from the condition that every person in the sample has the same probability of dying, and the probabilities are derived from the total number of deaths per year. However, the number of deaths changes yearly and also the population changes. We may handle one of the possible inaccuracies as follows. Suppose that 1200 persons per year die, so in ten years 12000 persons die. The probability that a certain person dies in ten years is estimated by  $12000/10^7$ . Thus, the probability that a specific person will not die in ten years is  $(1 - \frac{12}{10^4})$  (first two members of binomial expansion of  $(1 - \frac{12}{10^5})^{10}$ ). In total, we obtain the following estimate of the probability  $1 - (1 - \frac{12}{10^4})^{500} \doteq 0.4514$ . Observe that both estimates are very close one to each other.



We already met probabilities based on assumptions that other events have happened, we say “under certain hypothesis”. We talk about conditional probability, see 1.4.8. As we will see below, such estimates can be successfully derived also via Sage (which is also related to “Monte Carlo” computational method, cf. 1.4.10).

**1.D.11. Sage simulation.** What is the probability that when rolling two dice the sum is 7, if we know that neither of the rolls resulted in a 2? Estimate the solution via Sage.

**Solution.** Let  $B$  be the event that neither of the rolls results into 2, and let  $A$  be the event “sum is 7”. The set of all possible outcomes is again denoted by  $\Omega$ . Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{|A \cap B|}{|B|}.$$

Now, the number 7 can appear as a sum in four ways if there is no 2, that is,  $|A \cap B| = 4$ ,  $|B| = 5 \cdot 5 = 25$ . Thus

$$P(A|B) = \frac{4}{25} = 0.16.$$

Observe that  $P(A) = \frac{1}{6}$ , that is  $A$  and  $B$  are *not* stochastically independent.

Let us now run the simulation to estimate the resulting probability via Sage. Initially we use the parameter `num_rolls` and the method `random_element` (be careful that after copying-pasting the cells below, one needs to re-edit the functions to keep the spaces that appearing below, etc).

```
num_rolls = 100_000
die1=[IntegerRing().random_element(x=1,
    y=7) for _ in range(num_rolls)]
die2=[IntegerRing().random_element(x=1,
    y=7) for _ in range(num_rolls)]
```

– if we have two events  $A_1, A_2$  satisfying  $P(A_1 \cap A_2) > 0$ , then

$$P(A_1 \cap A_2) = P(A_2)P(A_1|A_2) = P(A_1)P(A_2|A_1).$$

All these numbers express (in a different manner) the probability that both events  $A_1$  and  $A_2$  occur. For instance, in the last case we first look whether the first event occurred. Then, assuming that the first has occurred, we look whether the second also occurs. Similarly, for three events  $A_1, A_2, A_3$  satisfying  $P(A_1 \cap A_2 \cap A_3) > 0$  we obtain

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2).$$

The probability that three events occur simultaneously can be computed as follows. Compute the probability that the first occurs, then compute the probability that the second occurs under the assumption that the first has occurred. Then compute the probability that the third occurs under the assumption that both the first and the second have occurred. Finally, multiply the results together.

In general, if we have  $k$  events  $A_1, \dots, A_k$  satisfying  $P(A_1 \cap \dots \cap A_k) > 0$ , then the theorem says

$$P(A_1 \cap \dots \cap A_k) = P(A_1)P(A_2|A_1) \cdots P(A_k|A_1 \cap \dots \cap A_{k-1}).$$

Notice that our condition that  $P(A_1 \cap \dots \cap A_k) > 0$  implies that all the hypotheses in the latter formula have got non-zero probabilities and thus all the conditional probabilities make sense. Indeed, each  $A_i$  is at least as big as the intersection and thus its probability is at least as big, see 1.4.3(7).

**1.4.9. Geometric probability.** In practical problems, the sample space may not be a finite set. The set  $\mathcal{A}$  of all events may not be the entire set of all subsets in  $\Omega$ . To generalise probability to such situations is beyond our scope now, but we can at least give a simple illustration.



Consider the plane  $\mathbb{R}^2$  of pairs of real numbers and a subset  $\Omega$  with known area  $|\Omega|$ . Events are represented by subsets  $A \subset \Omega$  (again with known areas). For the event set  $\mathcal{A}$  we consider some suitable system of subsets for which we can determine the area. An event  $A$  then occurs if a randomly chosen point from  $\Omega$  belongs to the subregion determined by  $A$ , otherwise the event does not occur.

Consider the problem of randomly choosing two numbers  $a < b$  in the interval  $[0, 1] \subset \mathbb{R}$ . All values  $a$  and  $b$  are chosen with equal probability. The question is “what is the probability that the interval  $(a, b)$  has length at least one half?” The choice of points  $(a, b)$  is actually the choice of a point  $[a, b]$  inside of the triangle  $\Omega$  with vertex points  $[0, 0]$ ,  $[0, 1]$ ,  $[1, 1]$  (see the diagram below).

We can imagine this as a description of a problem where a very tired guest at a party tries to divide a sausage with two cuts into three pieces for himself and his two friends. What is the probability that the middle part will be at least half of the sausage?

We select only rolls where neither dice resulted in a 2.

```
good_rolls = []
for x, y in zip(die1, die2):
    if x != 2 and y != 2:
        good_rolls.append((x, y))
```

Next we calculate the number of rolls with sum 7:

```
count_sum_7 = 0
for x, y in good_rolls:
    if x + y == 7:
        count_sum_7 += 1
```

The final probability is given by the ratio

```
result=count_sum_7/len(good_rolls)
print(n(result))
```

Sage’s output is 0.162572964225857 (this will slightly differ each time you run the code!), hence our estimate given by Sage has been close to the correct answer. □

**1.D.12.** Michael has two mailboxes, one at gmail.com and the other one at hotmail.com. His username is the same at both servers, but the passwords are different. He does not remember which password corresponds to which server. When typing in the password for accessing his mailbox, he makes a typo with probability 5% (i.e., if he tries to type in a specific password, he types what he intended with probability 95%). At the server hotmail.com, Michael typed in the username and a password, but the server told him that something is wrong. What is the probability that he chose the correct password but just mistyped? (we assume that the username is always typed correctly and that making a typo cannot turn a wrong password into a right one.)



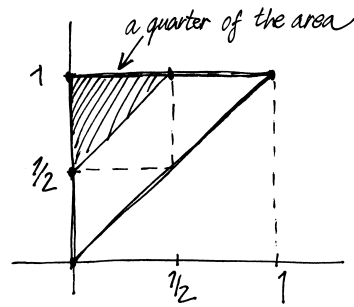
**Solution.** Let  $A$  be the event that Michael typed in a wrong password at hotmail.com. This event is the union of two disjoint events:

- $A_1$ : he wanted to type in the correct password and mistyped;
- $A_2$ : he wanted to type in the wrong password (the one from gmail.com) and either mistyped it or not.

We are looking for a conditional probability  $P(A_1|A)$  which according to the paragraph 1.4.8 is given by

$$\frac{P(A_1 \cap A)}{P(A)} = \frac{P(A_1)}{P(A_1 \cup A_2)} = \frac{P(A_1)}{P(A_1) + P(A_2)}.$$

Here, we have used the fact that  $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ , since  $A_1$  and  $A_2$  are disjoint. So, we only need to compute the probabilities  $P(A_1)$  and  $P(A_2)$ . The event  $A_1$  is the intersection of two independent events: Michael wanted to type in a correct password and Michael mistyped. According to the problem statement, the probability of the first event is  $1/2$  and the probability of the second event is  $1/20$ . In total, and since the events are independent, we get  $P(A_1) = \frac{1}{2} \cdot \frac{1}{20} = \frac{1}{40}$ . In addition,  $P(A_2) = \frac{1}{2}$ , and thus



Thus we need to determine the area of the subset which corresponds to points with  $b \geq a + \frac{1}{2}$ , that is, the interior of the triangle  $A$  bounded by the points  $[0, \frac{1}{2}]$ ,  $[0, 1]$ ,  $[\frac{1}{2}, 1]$ . We find  $P(A) = (1/8)/(1/2) = \frac{1}{4}$ .

Similarly, if we ask for the probability that some of the three guests will get at least half of the sausage, then we have to add the probabilities of two other events:  $B$  saying  $a \geq 1/2$  and  $C$  given as  $b \leq 1/2$ . Clearly they correspond to the lowest and the most right top triangles and thus they have got probabilities  $1/4$ , too. Thus the requested probability is  $3/4$ .

Equivalently we could have asked for the complementary event “all of them get less than a half” which clearly corresponds to the middle triangle and thus has probability  $1/4$ .

Try to answer on your own the question “what is the minimal prescribed length  $\ell$  such that the probability of choosing an interval  $(a, b)$  of length at least  $\ell$  is one half?”

**1.4.10. Monte Carlo methods.** One efficient method for computing approximate values is simulation by the relative occurrence of a chosen event.



We present an example. Let  $\Omega$  to be the unit square with vertices at  $[0, 0]$ ,  $[1, 0]$ ,  $[0, 1]$ , and  $[1, 1]$ . Let  $A$  be the intersection of  $\Omega$  with the unit disk centred at the origin. Then  $\text{area } A = \frac{1}{4}\pi$ . Suppose we have a reliable generator of random numbers  $a$  and  $b$  between zero and one. We then compute relative frequencies of how often  $a^2 + b^2 < 1$ . That is, that  $[a, b] \in A$ . Then the result (after a large number of attempts) should approximate the area of a quarter unit circle, that is  $\pi/4$  quite well. (Draw a picture yourselves!)

Of course, the well-known formula for the area of a circle with radius  $r$  is  $\pi r^2$ , where  $\pi = 3.14159\dots$ . It is an interesting question – why should the area of a circle be a constant multiple of the square of its radius? We will be able to prove this later. But experimentally, we can hint at this by the approach as above using squares of different sizes.

Numerical approaches based on such probabilistic principle are called *Monte Carlo methods*.

## 5. Plane geometry

So far we have been using elementary notions from the geometry of the real plane in an intuitive way. Now we will investigate in more detail how to deal with the need to describe “position in the plane” and to find some relation between positions of distinct points in the plane.



$P(A) = P(A_1) + P(A_2) = \frac{1}{40} + \frac{1}{2} = \frac{21}{40}$ . This gives

$$P(A_1|A) = \frac{P(A_1)}{P(A)} = \frac{\frac{1}{40}}{\frac{21}{40}} = \frac{1}{21}.$$

With the Sage cell

```
PA1=1/2*1/20; PA2=1/2; PA=PA1+PA2
PA1condA=PA1/PA
print("P(A1|A)=", PA1condA)
```

we obtain the verification  $P(A_1|A) = 1/21$ . □

**1.D.13.** Consider a deck of 32 cards. If we draw twice one card, what is the probability that the second drawn card is an ace, if we return the first card; and when we don't return the first card (then there are 31 cards in the deck).

**Solution.** If we return the card in the deck, we are just repeating the experiment, which has 32 possible results (having the same probability), and exactly four of them are favourable. Thus we see that the probability is  $p = 1/8$ . In the second case when we do not return the card, the probability is the same. Indeed, notice that when drawing all the cards one by one, the probability of an ace as the first card is identical to the probability getting an ace in the second draw.

We can also apply the conditional probability, splitting the event into two disjoint options of drawing or not drawing an ace as the first card:  $p = \frac{4}{32} \cdot \frac{3}{31} + \frac{28}{32} \cdot \frac{4}{31} = \frac{1}{8}$ . □

*The concept of geometric probability is useful if we can identify individual events with positions on a line, in a plane, in a space, etc., while the sample space  $\Omega$  is a region with known length, area, volume, respectively. The favorable positions are then represented by measurable objects, too. In analogy to the classical probability, we define the probability as the ratios of length, areas, volumes, etc., to those of  $\Omega$ , see ???. This reflects the idea that all subregions of the same measure are hit with the same probability. We present one example (and we include more of them in the final section of this chapter, starting at page 61).*

**1.D.14.** In a certain country, a bus departs from town A to town B once a day at a random time between eight a.m. and eight p.m. Once a day in the same time interval another bus departs in the opposite direction. The trip in either direction takes five hours. Compute the probability that the buses meet, assuming they use the same route.

**Solution.** The sample space is a square  $12 \times 12$ . If we denote the time of the departure of the buses as  $x$  and  $y$  respectively, then they meet on the trail if and only if  $|x - y| \leq 5$ . This inequality determines the region in the square of "favourable events". This is a complement to the union of two right-angled isosceles triangles with legs of length 7, see also the figure below.

Our tools will be mappings. We will mainly consider mappings  $F$  which, to (ordered) pairs of values  $(x, y)$ , assign pairs  $(w, z) = F(x, y)$ . Such a mapping will consist of two functions  $w(x, y)$  and  $z(x, y)$ , each depending on two arguments  $x, y$ . This will also serve as a gentle introduction to the part of mathematics called *Linear algebra*, with which we will deal in the subsequent three chapters.

**1.5.1. Vector space  $\mathbb{R}^2$ .** We view the "plane" as a set of pairs of real numbers  $(x, y) \in \mathbb{R}^2$ . We will call these pairs *vectors* in  $\mathbb{R}^2$ . For such vectors we can define addition "coordinate-wise", that is, for vectors  $u = (x, y)$  and  $v = (x', y')$  we set

$$u + v = (x + x', y + y').$$

Since all the properties of commutative groups hold for individual coordinates, these hold for our new vector addition too. In particular there exists a *zero vector*  $0 = (0, 0)$ , such that  $v + 0 = v$ . We use the same symbol 0 for the vector and for the number zero on purpose. The context will always make it clear which "zero" it is.

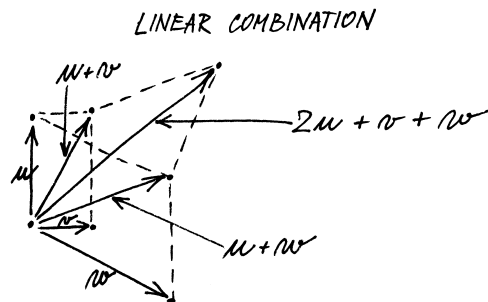
Next we define scalar multiplication of vectors. For  $a \in \mathbb{R}$  and  $u = (x, y) \in \mathbb{R}^2$ , we set

$$a \cdot u = (ax, ay).$$

Usually we will omit the symbol  $\cdot$  and use the juxtaposition of the symbols  $av$  to denote the scalar multiple of a vector.

We can directly check other properties for scalar multiplication by  $a$  or  $b$  and addition of vectors  $u$  and  $v$ . For instance  $a(u+v) = au + av$ ,  $(a+b)u = au + bu$ ,  $a(bu) = (ab)u$ .

We use the same symbol  $+$  for both vector addition and scalar addition.



Now we take a very important step. Define vectors  $e_1 = (1, 0)$  and  $e_2 = (0, 1)$ . Every vector can then be written uniquely as

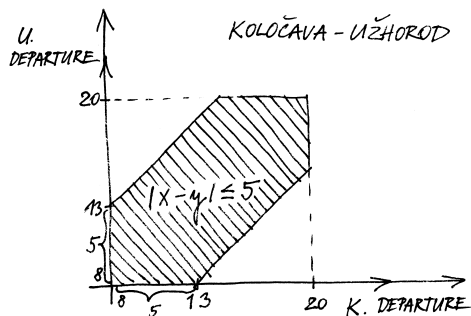


$$u = (x, y) = x e_1 + y e_2.$$

The expression on the right is called a *linear combination of vectors  $e_1$  and  $e_2$*  (with coefficients  $x$  and  $y$ ). The pair of vectors  $\underline{e} = (e_1, e_2)$  is called the *standard basis* of the vector space  $\mathbb{R}^2$ .

As shown on the picture, these operations are easy to imagine if we consider the vectors  $v$  to be arrows starting at the origin  $0 = (0, 0)$  and ending at the position  $(x, y)$  in the plane.





Its area in total is 49, so the area of the “favourable part” is  $144 - 49 = 95$ . The probability is  $p = \frac{95}{144} \doteq 0.66$ .  $\square$

### E. Plane geometry

*In this section we handle elementary geometric objects in the real plane  $\mathbb{R}^2$ , including points, vectors, lines, circles, and polygons. Our explanation will be supported by Sage. We begin with a series of standard tasks related to the fundamental notion of lines. Recall that a line in the plane is given “implicitly” by the equation  $ax + by = c$  (the line is the set of all points  $[x, y]$  given by the solutions with fixed parameters  $a, b, c$ ). We shall mostly write  $\ell, p, q$  for lines. Lines can be also “parametrized” using equations of the form  $x = x_0 + tv_1$  and  $y = y_0 + tv_2$ , where  $P = [x_0, y_0]$  is a point on the line,  $(v_1, v_2)$  is the direction vector, and  $t \in \mathbb{R}$  is a parameter, see 1.5.3 for more details.*



**1.E.1. Points versus vectors.** Points in the plane will be denoted with capital letters, as  $P = [x, y]$ , where  $x, y \in \mathbb{R}$  are the coordinates, while the vectors shifting  $P$  to  $Q = [x + u_1, y + u_2]$  will be written as  $u = (u_1, u_2)$ . This follows the needs to distinguish the “position in the plane”  $P$  and the “change of the position”  $u$ . Notice, the same notation is used for closed and open intervals on a real line, cf. Chapter 5, but we do not see much space for confusion here. The reader is advised to look carefully at the explanation starting in 1.5.2, to fully appreciate working with both concepts as couples of reals, as soon as we fix our coordinates. This approach will be crucial when dealing with affine and Euclidean geometries in Chapter 4.

We shall also adopt the matrix calculus which brings the notation for both points and vectors in the form of columns, i.e., instead of  $u = (x, y)$  or  $P = [x, y]$ , we often write  $\begin{pmatrix} x \\ y \end{pmatrix}$ .

**1.E.2.** Determine the line  $\ell$  given by  $\{x = 2 - t, y = 1 + 3t : t \in \mathbb{R}\}$  implicitly. Next present an illustration of  $\ell$  using Sage for those  $x$  with  $-5 \leq x \leq 5$ .

**Solution.** We need to find the implicit representation of  $\ell$ , and this can be done by eliminating the parameter  $t$ . Solving the first equation with respect to  $t$  gives  $t = 2 - x$ . Replacing this in the second equation gives  $y = 1 + 3(2 - x)$ , that is  $y + 3x - 7 = 0$ , which is the solution. To present a line via Sage, there are many alternatives. Here we use the implicit form  $y = -3x + 7$ , and then the code can be

The addition of two such arrows is then given by the parallelogram law: Given two arrows starting at the origin, their sum is the arrow given by the diagonal arrow (also starting at the origin) of the parallelogram with the two given arrows as adjacent sides. Multiplication by a scalar  $a$  corresponds to stretching the arrow to its  $a$ -multiple. This includes negative scalars, where the direction of the vector is reversed.

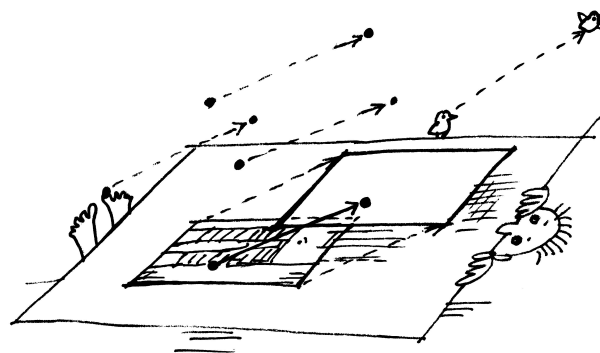
If we choose any two non-zero vectors  $u, v$  such that neither of them is a multiple of the other, then they form a basis of  $\mathbb{R}^2$ , too, i.e. each other vector can be (uniquely) written as a linear combination of these. This seems to be clear from the picture, and we may easily write down the two equations for the coefficients and to solve them (as we shall deduce in a while).

**1.5.2. Points in the plane.** In geometry, we should distinguish between the points in the plane (as for instance the chosen origin  $O$  above), and the vectors as the arrows describing the difference between two such points. We will work in fixed standard coordinates, that is, with pairs of real numbers, but for better understanding we will now distinguish vectors written in parentheses and denoted for a moment by bold face letters like  $\mathbf{u}, \mathbf{v}$ , instead of brackets (which we use for coordinates of points in the plane). Points themselves are denoted by capital latin letters.

Even if we view the entire plane as pairs of real numbers in  $\mathbb{R}^2$ , we may understand adding two such couples as follows. The first couple of coordinates describes a point  $P = [x, y]$ , while the other one denotes a vector  $\mathbf{u} = (u_1, u_2)$ . Their sum  $P + \mathbf{u}$  corresponds to adding the (arrow) vector  $\mathbf{u}$  to the point  $P$ . If we fix the vector  $\mathbf{u}$ , we call the resulting mapping

$$P = [x, y] \mapsto P + \mathbf{u} = [x + u_1, y + u_2]$$

the *shift of the plane* (or *translation*) by the vector  $\mathbf{u}$ .

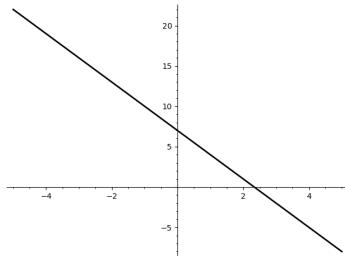


Thus, the vectors in  $\mathbb{R}^2$  can be understood in more abstract way as the shifts in the plane (sometimes called the *free vectors* in elementary geometry texts).

The standard coordinates on  $\mathbb{R}^2$ , understood as pairs of real numbers are not the only ones. We can put a coordinate system on the plane with our choosing.

```
e11=-3*x+7
plot(e11, -5, 5, color="black", thickness=2)
```

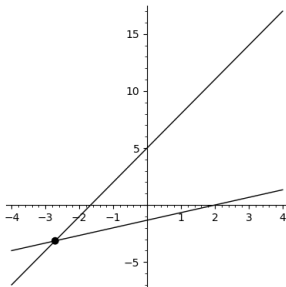
Or simply type `e11 = -3 * x + 7; plot(e11, -5, 5)`, or `plot(-3 * x + 7, -5, 5)`, without the options for the colour and the thickness of the line. The figure that Sage returns via the very first choice is given here:



□

**1.E.3.** Consider the lines  $p, q$  given in the form  $[2, 0] + t(3, 2)$ , and  $[-1, 2] + s(1, 3)$ , with  $t, s \in \mathbb{R}$ , respectively. Present them in implicit form, and next examine if they intersect. In positive case, determine explicitly the (unique) intersection point. Verify your answer via Sage.

**Solution.** The coordinates of the points on the first line are given by the parametric equations  $\{x = 2 + 3t, y = 0 + 2t\}$ . Eliminating  $t$  from the equations we obtain the implicit form of  $p$ :  $2x - 3y - 4 = 0$ . Similarly, the points on the line  $q$  are given by  $\{x = -1 + s, y = 2 + 3s\}$ .



By eliminating  $s$ , we get the implicit form of  $q$ , namely  $3x - y + 5 = 0$ . The unique solution  $(x^*, y^*)$  of the system of the equations  $\{2x - 3y - 4 = 0, 3x - y + 5 = 0\}$ , if any, determines the coordinates of the unique intersection point  $P = (x^*, y^*)$  of  $p, q$ , if any. We compute  $P = [-\frac{19}{7}, -\frac{22}{7}]$ . As we have seen, Sage offers the command `solve`. Here its application comes with the cell

```
x, y=var("x, y")
eq1=2*x-3*y-4; eq2=3*x-y+5
solve([eq1==0, eq2==0], x, y)
```

which returns  $[[x == (-19/7), y == (-22/7)]]$ . As for the given plot that joins both lines, we have applied the Sage cell

COORDINATES IN THE PLANE  $\mathbb{R}^2$

Choose any point in the plane, and call it the origin  $O$ . All other points  $P$  in the plane can be identified with the vectors (arrows)  $\overrightarrow{OP}$  with their tails at the origin.

Choose any point other than  $O$  and call it  $E_1$ . This defines the vector  $e_1 = \overrightarrow{OE_1} = (1, 0)$ . Choose any other point  $E_2$  so that  $O, E_1, E_2$  are distinct and not collinear. This defines the vector  $e_2 = \overrightarrow{OE_2} = (0, 1)$ .

Then every point  $P = (a, b)$  in the plane can be described uniquely as  $P = O + a e_1 + b e_2$  for real  $a, b$ , or in vector notation,  $\overrightarrow{OP} = a e_1 + b e_2$ .

Translation, by adding a fixed vector, can be used either to shift the coordinate system (including the origin), or to shift sets of points in the plane. Notice that the vector corresponding to the shift of the point  $P$  into the point  $Q$  is given as the difference  $Q - P$  (in any coordinates). Thus we shall also use this notation for the vector  $\overrightarrow{PQ} = Q - P$ .

For each choice of coordinates, we have two distinct lines for the two axes. The origin is the point of intersection. Other way round, each choice of two non-parallel lines, together with the scales on each of them defines coordinates in the plane. They are called *affine coordinates*.

Clearly each nontrivial triangle in the plane with vertices  $O, E_1, E_2$  defines coordinates where this triangle is defined by points  $[0, 0], [1, 0], [0, 1]$ . Thus we may say that in the geometry of plane, "all nontrivial triangles are the same, up to a choice of coordinates".

**1.5.3. Lines in the plane.** Every line is parallel to a (unique) line through the origin. To define a line, we therefore need two ingredients. One is a non-zero vector which describes the direction of the line. Call it  $\mathbf{v} = (v_1, v_2)$ . The other is a point  $P_0 = [x_0, y_0]$  on the line. Every point on the line is then of the form



$$P(t) = P_0 + t\mathbf{v}, \quad t \in \mathbb{R}.$$

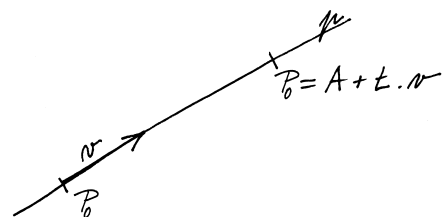
PARAMETRIC DESCRIPTION OF A LINE

We may understand the line  $p$  as the set of all multiples of the vector  $v$ , shifted by the vector  $(x_0, y_0)$ . This is called the *parametric description* of the line:

$$p = \{P \in \mathbb{R}^2; P = P_0 + t\mathbf{v}, t \in \mathbb{R}\}.$$

The vector  $v$  is called *direction vector* of the line  $p$ .

LINE EQUATION



```
z=var("z")
y1=(2/3)*z-(4/3); y2=3*z+5
P = plot([],figsize=(4,4))
P += plot(y1,z,(-4,4), color="black")
P += plot(y2,z,(-4,4), color="black")
P += point((-19/7, -22/7), size=50,
           color="black"); show(P)
```

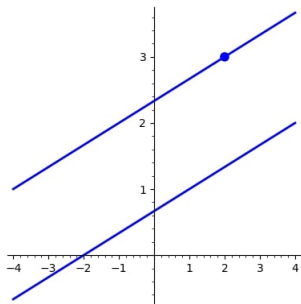
**1.E.4. Remark.** In 1.E.3 you can also solve the problem by substituting the parametric equations of the line  $q$  into the implicit equation of  $p$ , which gives  $2(-1+s)-3(2+3s)-4 = 0$ . This equation has the unique solution  $s = -12/7$ , hence returning back to the parametric equation of  $q$  we obtain the coordinates of  $P$  as previously mentioned.

In Sage you can directly plot a line given in parametric form using the `parametric_plot` command. For example, to plot the lines from 1.E.3 you can use the following Sage cell (check the resulting figure in your editor). Sage can also find the intersection point by solving the systems of the two equations for the parameters  $s, t$  - try it out yourselves!

```
t = var("t")
P1= parametric_plot((2+3*t, 2*t), (t, -3, 3),
                   color="blue", legend_label="line p")
P2= parametric_plot((-1+t, 2+3*t), (t, -3,
3), color="black", legend_label="line q")
Q= point((-19/7, -22/7), size=70)
PL=P1+P2+Q; show(PL)
```

**1.E.5.** Determine the intersection of the lines  $x + y - 4 = 0$  and  $\{x = -1 + 2t, y = 2 + t : t \in \mathbb{R}\}$ .

**1.E.6.** Find the line  $p$  which passes through the point  $P = [2, 3] \in \mathbb{R}^2$  and is parallel to line  $q$  given by  $x - 3y + 2 = 0$ . Provide a figure in Sage illustrating the situation.



**Solution.** On the real plane for two parallel lines there are two chances: either they overlap, or they have no intersection. Since the line  $p$  should pass from the point  $[2, 3]$  and this point does not belong to  $q$  (since it does not satisfy its equation), we deduce that  $p, q$  have no intersection.

This means that  $p$  is of the form  $x - 3y + c = 0$ , for some  $c \in \mathbb{R}$ , with  $c \neq 2$ . To find  $c$ , we substitute the coordinates of the point  $P$  into this equation. It reveals  $c = 7$ , which means that  $p$  is implicitly given by  $y = \frac{1}{3}x + \frac{7}{3}$ . For the figure presented above, we have used the cell

```
x=var("x")
y1=(1/3)*x+(2/3)
y2=(1/3)*x+(7/3)
a = plot([],figsize=(4,4))
```

In the chosen coordinates, the point  $P(t) = [x(t), y(t)]$  is given as

$$x = x(t) = x_0 + t v_1, \quad y = y(t) = y_0 + t v_2$$

We can eliminate  $t$  from these two equations to obtain

$$-v_2x + v_1y = -v_2x_0 + v_1y_0.$$

Since the vector  $v = (v_1, v_2)$  is non-zero, at least one of the numbers  $v_1, v_2$  is non-zero. If one of the coordinates  $v_1$  or  $v_2$  is zero, then the line is parallel to one of the coordinate axis.

#### IMPLICIT DESCRIPTION OF A LINE

The general equation of the line in the plane is

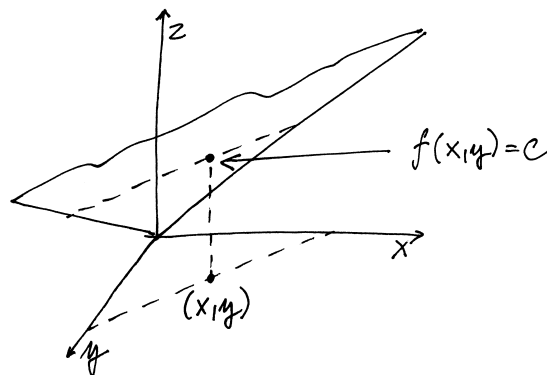
$$(1) \quad ax + by = c,$$

with  $a$  and  $b$  not both zero. The relation between the pair of numbers  $(a, b)$  and the direction vector of the line  $\mathbf{v} = (v_1, v_2)$  is

$$(2) \quad av_1 + bv_2 = 0.$$

We can view the left hand side of the equation (1) as a function  $z = f(x, y)$  mapping each point  $[x, y]$  of the plane to a scalar and the line corresponds to the prescribed constant value of this function. We shall see soon that (2) says the vector  $(a, b)$  is perpendicular to the direction of the line.

#### GRAPH OF A FUNCTION $f(x, y)$



Suppose we have two lines  $p$  and  $q$ . We ask about their intersection  $p \cap q$ . That is a point  $[x, y]$  which satisfies the equations of both lines simultaneously. We write them as

$$(3) \quad \begin{aligned} ax + by &= r \\ cx + dy &= s. \end{aligned}$$

Again, we can view the left side as a mapping  $F$ , which to every pair of coordinates  $[x, y]$  of the point  $P$  in the plane assigns a vector of values of two scalar functions  $f_1$  and  $f_2$  given by the left sides of the particular equations (3). Hence we can write our two scalar equations as one vector equation  $F(\mathbf{v}) = \mathbf{w}$ , where  $\mathbf{v} = (x, y)$  and  $\mathbf{w} = (r, s)$ .

Notice that the two lines are not parallel if and only if they have a unique point in their intersection.

```
a += plot(y1, x, (-4, 4), thickness=2)
a += plot(y2, x, (-4, 4), thickness=2)
a += point((2, 3), size=70);show(a)
```

Observe that in this case we asked a different size for the thickness of lines, in comparison with the previous figures.  $\square$

**1.E.7.** Consider the following lines in  $\mathbb{R}^2$ :

$$\begin{aligned} p_1 : 2x + 3y - 4 &= 0, & p_2 : x - y + 3 &= 0, \\ p_3 : -2x + 2y &= -6, & p_4 : -x - \frac{3}{2}y + 2 &= 0, \\ p_5 : x = 2 + t, & y = -2 - t, & t \in \mathbb{R} \end{aligned}$$

Determine which lines are parallel to each other.  $\circ$

**1.E.8.** Find a parametric equation of the line  $\ell$  which goes through the points  $P_1 = [1, 3]$  and  $P_2 = [-2, 1]$  on  $\mathbb{R}^2$ . What is the corresponding implicit equation of  $\ell$ ?

**Solution.** The parametric equation of a line  $\ell$  on the Euclidean plane passing through two points  $P_1 = [u_1, v_1]$  and  $P_2 = [u_2, v_2]$  on  $\mathbb{R}^2$  is given by

$$P_1 + t\overrightarrow{P_1P_2} = P_1 + t(P_2 - P_1),$$

where  $\overrightarrow{P_1P_2} := P_2 - P_1 = (u_2 - u_1, v_2 - v_1)$  is the displacement vector from  $P_1$  to  $P_2$ . Applying this rule for the given points, we get  $P_2 - P_1 = (-3, -2)$  and so the parametric equation of  $\ell$  is given by  $[1, 3] + t(-3, -2) = [1, 3] - t(3, 2)$ , with  $t \in \mathbb{R}$ . Hence we have  $x = 1 - 3t$  and  $y = 3 - 2t$ . By eliminating  $t$  we get  $y = \frac{2}{3}x + \frac{7}{3}$ . As a verification one can check that both the points  $P_1, P_2$  satisfy the equation of  $\ell$ .  $\square$

**1.E.9.** Planar soccer player shoots a ball from the point  $F = [1, 0]$  in the direction of the vector  $u = (3, 4)$  hoping to hit the goal which is a line segment from the point  $A = [23, 36]$  to  $B = [26, 30]$ . Does the ball fly towards the goal?

**Solution.** The ball travels along the line  $[1, 0] + t(3, 4)$ . The line segment joining  $A$  and  $B$ , has the parametrization  $[23, 36] + s(3, -6)$ . The intersection of these lines is given by equations  $1 + 3t = 23 + 3s$  and  $4t = 36 - 6s$ , with the solution  $t = 8, s = 2/3$ . As  $0 < 2/3 < 1$  the intersection is in the line segment between  $A$  and  $B$ , and so the ball hits the goal.

Another approach (closed to the function of our vision) is based on the slopes of the vectors  $\overrightarrow{FA}, u = (3, 4)$ , and  $\overrightarrow{FB}$  (i.e. the ratio of the coordinates, we shall come back to that later). Since  $\frac{36}{22} > \frac{4}{3} > \frac{30}{25}$ , the player scores.  $\square$

*Next, we shall touch the basic computational tools in the entire Mathematics, the matrix calculus. For now, we shall restrict our attention to tasks in the plain only, and the reader can find introductory concepts from the so called "linear algebra" in 1.5.4, also restricted to the real plane  $\mathbb{R}^2$ . Below we will first experience the operations of addition and multiplication of matrices, and exercise how to compute with them more or less exactly as with scalars. We shall return to the geometry of the plane with the new tools at hand later.*

**1.5.4. Linear mappings and matrices.** Mappings  $F$  with



which we have worked with when describing intersection of lines have one very important property in common: they preserve the operations of addition and multiplication with vectors and scalars, that is they preserve linear combinations:

$$F(a \cdot \mathbf{v} + b \cdot \mathbf{w}) = a \cdot F(\mathbf{v}) + b \cdot F(\mathbf{w})$$

for all  $a, b \in \mathbb{R}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^2$ . We say that  $F$  is a *linear mapping* from  $\mathbb{R}^2$  to  $\mathbb{R}^2$ , and write  $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ . This can be also described with words — linear combination of vectors maps to the same linear combination of their images, that is linear mappings are those mappings which preserve linear combinations.

We have already encountered the same behaviour in the equation 1.5.3(1) for the line, where the linear mapping in question was  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  and its prescribed value  $c$ . That is also the reason why the values of the mapping  $z = f(x, y)$  are on the image depicted as a plane in  $\mathbb{R}^3$ .

We can write such a mapping using *matrices*. By a matrix we mean a rectangular array of numbers, for instance

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{or} \quad \mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

We speak of a (square  $2 \times 2$ ) matrix  $A$  and (column) vector  $\mathbf{v}$ . Multiplication of matrices, row by column, is defined as follows:

$$A \cdot \mathbf{v} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}.$$

We introduce some more tools for vectors and matrices. Our goal is to compute with matrices in a similar way as we do it with scalars.

We define the product  $C = A \cdot B$  of two square matrices  $A$  and  $B$  applying the above formulas to individual columns of the matrix  $B$  and writing the resulting column vectors again as columns in the matrix  $C$ .

In order to multiply two vectors  $\mathbf{w}$  and  $\mathbf{v}$  in a similar way, we can write the vector  $\mathbf{w}$  as a row of numbers (the transposed vector)  $\mathbf{w}^T$ . Then the product of  $\mathbf{w}^T$  and  $\mathbf{v}$  is



$$\mathbf{w}^T \cdot \mathbf{v} = (r \quad s) \cdot \begin{pmatrix} x \\ y \end{pmatrix} = rx + sy.$$

This is the *scalar product of vectors*  $\mathbf{v}$  and  $\mathbf{w}$  introduced a while ago.

We can easily check the associativity of multiplication (do it for general matrices  $A, B$  and a vector  $\mathbf{v}$  in detail):

$$(A \cdot B) \cdot \mathbf{v} = A \cdot (B \cdot \mathbf{v}).$$

Instead of a vector  $\mathbf{v}$  we can write any  $(2 \times 2)$  matrix  $C$ . In a similar way, distributivity also holds:

$$\begin{aligned} A \cdot (B + C) &= A \cdot B + A \cdot C \\ (A + B) \cdot C &= A \cdot C + B \cdot C. \end{aligned}$$

**1.E.10. Quick comments on notation.** In order to simplify our notation, for the multiplication of two matrices  $A, B$  (of appropriate size) we just write  $AB$ , instead of  $A \cdot B$  (which is the complete notation we use in the other column, but only in this chapter). Later on we shall use the operator “ $\cdot$ ” for the scalar product  $u \cdot v$  between two vectors  $u, v$  (which we do in the other column as well).



In the sequel, we shall write  $\text{Mat}_{m,n}(\mathbb{R})$  for the set of all  $m \times n$  real matrices (so in this section both  $m$  and  $n$  will be either 1 or 2). For  $m = n$ , i.e., square matrices, we will use the notation  $\text{Mat}_n(\mathbb{R})$ . If  $m = 1, n = 2$ , the object is a couple of scalars written in a row, if  $n = 1$  and  $m = 2$ , we deal with a vector written as a column (while  $\text{Mat}_1(\mathbb{R})$  coincides with  $\mathbb{R}$ ). Moreover, for  $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in \text{Mat}_2(\mathbb{R})$  we denote by  $\det(A)$  and  $\text{tr}(A)$  the determinant and trace of  $A$ , which are the real numbers defined by  $\det(A) = \alpha\delta - \beta\gamma$  and  $\text{tr}(A) = \alpha + \delta$ , respectively. On all matrices, there is the transposition operation,  $A \mapsto A^T$  swapping rows and columns, and the product by scalars  $cA$ , for  $c \in \mathbb{R}$ . In Chapter 2, we will extend the matrix calculus to general positive integers  $m$  and  $n$ , and general scalars.

**1.E.11.** Let  $A, B$  be two real square matrices. How many arithmetic operations do we need to compute the product  $AB$ ?

**Solution.** Given  $A, B \in \text{Mat}_2(\mathbb{R})$ , each entry of the product  $AB$  requires 2 multiplications and  $2 - 1 = 1$  addition. Since there are  $2^2 = 4$  entries in  $AB$ , we need  $4(2 + 1) = 12$  operations in total.

We may notice, that the same operation of “multiplying” rows and columns for general  $A, B \in \text{Mat}_n(\mathbb{R})$  requires  $n$  multiplications and  $n - 1$  additions for each term in the result. Thus, we shall see in the next chapter that in total we need  $n^2(n+n-1) = n^2(2n-1)$  arithmetic operations to compute  $AB$  for two real square matrices  $A, B$ .<sup>8</sup> □

**1.E.12.** Compute  $v = 2(A - B)^T C u$ , for  $A = \begin{pmatrix} 0 & 5 \\ -2 & 2 \end{pmatrix}$ ,  $B = \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix}$ ,  $C = \begin{pmatrix} 2 & -2 \\ 4 & 5 \end{pmatrix}$ , and  $u = (3, 2)^T$ , respectively. ○

**1.E.13. Matrices in Sage.** Our aim is to compute with matrices and vectors just like with scalars. Sage can do this. Let us look at simple computations like in 1.E.12. We have to tell Sage to use the objects of the right class. The commands `matrix` and `vector` do that for us. Let us type



```
A = matrix([[0, 5], [-2, 2]])
B = matrix([[2, 0], [-1, 1]])
C = matrix([[2, -2], [4, 5]])
u = vector([3, 2])
v = 2*transpose(A-B)*C*u; print(v)
```

<sup>8</sup>This means, that the complexity of matrix multiplication is “polynomial”, essentially of the size of  $n^3$ . With big matrices, this is an essential issue in theoretical computer science to make the power lower. Volker Strassen introduced his famous algorithm reaching the complexity of the size  $n^{2.807}$  in 1969, and as of July 2023, the fastest known algorithm performs with the complexity size of  $n^{2.371866}$ .

But the commutativity does not hold. For example,

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

This last product also shows the existence of divisors of zero.

Notice that the mapping defined by multiplication of vectors with a fixed matrix is a linear mapping, i.e. it respects linear combinations. On the other hand, every linear mapping  $F$  has to be completely given by its values on the two vectors  $e_1 = (1, 0)$  and  $e_2 = (0, 1)$  in the standard basis and these values appear in the columns of the matrix  $A$ , which expresses  $F$  as matrix multiplication (check this observation carefully!).

With matrices and vectors we can write the equations for lines and points respectively as

$$\mathbf{u}^T \cdot \mathbf{v} = (a \ b) \cdot \begin{pmatrix} x \\ y \end{pmatrix} = c$$

$$A \cdot \mathbf{v} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \\ s \end{pmatrix} = \mathbf{w}.$$

**1.5.5. Determinant of matrix.** The procedure of finding the intersection of lines described in 1.5.3 fails in some special cases. For instance the intersection of two parallel lines is either empty (when the lines are parallel but distinct) or the line itself (when the lines are identical). This condition occurs when the ratios  $a/c$  and  $b/d$  are the same, that is

$$(1) \quad ad - bc = 0.$$

Note that this expression already takes care of the cases, where either  $c$  or  $d$  is zero.

The expression on the left in (1) is called the *determinant* of the matrix  $A$ . We write it as

$$\det A = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

Our discussion can be now expressed as follows:

**Proposition.** *The determinant is a real valued function  $\det A$  defined for all square  $2 \times 2$  matrices  $A$ . The (vector) equation  $A \cdot \mathbf{v} = \mathbf{u}$  has a unique solution for  $\mathbf{v}$  if and only if  $\det A \neq 0$ .*

So far, we have worked with pairs of real numbers in the plane. Equally well we might pose exactly the same questions for points with integer coordinates and lines with equations with integer coefficients. Notice that the latter requirement is equivalent to considering rational coefficients in the equations. We have to be careful which properties of the scalars we exploit.

In fact, we needed all the properties of the field of scalars when discussing the solvability of the system of two equations — try to think it through. At least, we can be sure that the intersection of two non-parallel lines with rational coefficients is a point with rational coefficients again. The case of integer coefficients and coordinates is more difficult. We shall come

Notice, Sage does all the multiplications properly (including multiplying correctly all the components by 2), and returns the components  $(-52, 64)$  of the vector  $v$ . The methods of addition and multiplication are appropriately used for the objects, with the same symbols as with scalars. Of course, we may introduce a matrix with other scalar type of entries. For example, the real float type entries are invoked by  $A = \text{matrix}(\text{RR}, [[0, 5], [-2, 2]])$  (recall that  $\text{RR}$  indicates the field of float type real numbers in Sage), and similarly with  $\text{CC}$  for complex float type. Sage also offers the methods  $\text{transpose}(A)$  and  $\text{determinant}(A)$  (or write  $A.\text{transpose}()$  and  $A.\text{determinant}()$ , respectively), which return the transpose and determinant of a matrix  $A$ . Further methods and comments will appear when using Sage below.

**1.E.14.** Prove the statements listed below, where  $A, B, C$  are the matrices from 1.E.12:

- (a)  $(A - B)^2 \neq A^2 - 2AB + B^2$ ;
- (b)  $D := 2ABC - BCA - CAB \neq 0$ ;<sup>9</sup>
- (c)  $\frac{c(AB) - (c(AB))^T}{\sqrt{2}} = \frac{11\sqrt{2}c}{2} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ , with  $c \neq 0$ ;
- (d)  $\det(AB) - \det(A)\det(B) = 0$ ;
- (e) The determinant of the matrix

$$F := \det(A)B - \det(B)A - \text{tr}(C)C$$

is an integer divisible by the numbers 1, 2, 19 and 38.

**1.E.15.** Give an example of matrices  $A$  and  $B$  for which

- (a)  $(A + B)(A - B) \neq A^2 - B^2$ .
- (b)  $(A + B)(A + B) \neq A^2 + 2AB + B^2$ .



The so called “affine geometry” of the real plane is characterized by the concept of “going straight with constant velocity” from a given point, i.e., along a line. The mappings  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  of the plane which preserve this feature are called “affine”. Those fixing the origin  $O = [0, 0]$  are just the “linear ones”,

$$F(au + bv) = aF(u) + bF(v),$$

for all  $a, b \in \mathbb{R}$  and  $u, v \in \mathbb{R}^2$ . Remarkably, all affine maps are obtained by additionally allowing the “shifts”  $P \mapsto P + u$ ,  $P \in \mathbb{R}^2$ , for some fixed vectors  $u$ .

In coordinates, the linear mappings are given by matrix multiplication, while the affine mapping include also addition of constant vectors. See the detailed explanation in the other column, e.g., 1.5.6. Due to the associativity of matrix multiplication, composition of mappings corresponds to multiplication of matrices, too. Moreover, invertible affine mappings are also understood as changes of coordinates, see 1.5.6, thus we call them affine transformations.

<sup>9</sup>Here we denote by  $0$  the zero  $2 \times 2$  matrix.

back to this in the next chapter. In particular we shall see that the equation 1.5.3(3) with fixed integer coefficients  $a, b, c, d$  has a unique integer solution for all integer values  $(r, s)$  if and only if the determinant is  $\pm 1$ .

**1.5.6. Affine mappings.** We now investigate how the matrix notation allows us to work with simple mappings in the affine plane. We have seen that matrix multiplication defines a linear mapping. Shifting  $\mathbb{R}^2$  by a fixed vector  $\mathbf{w} = (r, s) \in \mathbb{R}^2$  in the affine plane can be also easily written in matrix notation:



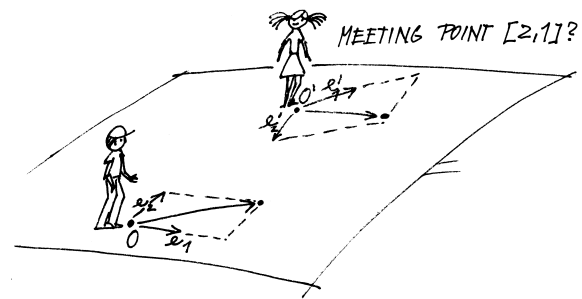
$$P = \begin{pmatrix} x \\ y \end{pmatrix} \mapsto P + \mathbf{w} = \begin{pmatrix} x + r \\ y + s \end{pmatrix}.$$

If we add a fixed vector to the result of a linear mapping then we have the expression

$$\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} \mapsto A \cdot \mathbf{v} + \mathbf{w} = \begin{pmatrix} ax + by + r \\ cx + dy + s \end{pmatrix}.$$

In this way we have described all affine mappings of the plane to itself.

Such mappings allow us to recompute coordinates which arise by different choices of origins or bases. We shall come back to this in detail later.



**1.5.7. The distance and angle.** Now we consider distance. We define the length of the vector  $\mathbf{v} = (x, y)$  to be

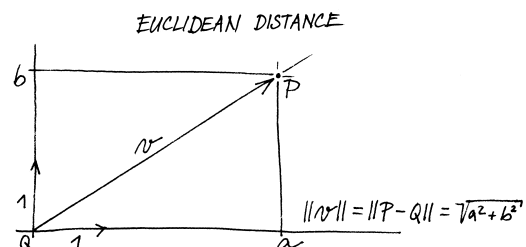
$$\|\mathbf{v}\| = \sqrt{x^2 + y^2}.$$

Immediately we can define notions of distance, angle and rotation in the plane.

DISTANCE IN THE PLANE

The distance between the points  $P, Q$  in the plane is given as the length of the vector  $\overrightarrow{PQ}$ , i.e.  $\|Q - P\|$ . Obviously, the distance does not depend on the ordering of  $P$  and  $Q$  and it is invariant under shifts of the plane by any fixed vector  $\mathbf{w}$ .

The Euclidean plane is an affine plane with distance defined as given above.



**1.E.16.** Decide whether the mappings  $F, G : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  given by

$$F : (x, y) \mapsto (7x - 3y, -2x + 5y),$$

$$G : (x, y) \mapsto (2x + 2y - 4, 4x - 9y + 3),$$

are linear. Write them via matrix multiplication, analyze the meaning of the individual columns, and check that their composition is expressed via matrix multiplication, too.  $\circ$

*Affine transformations play a crucial role in computer graphics and computer vision. Computer graphics involve creating and manipulating images, displayed or animated on a computer screen. Modern computer-aided design (CAD) programs and tools like LibreCAD have a wide range of applications, from designing satellites to creating computer games and educational software. Thus these tools are particularly valuable for engineers and designers working in various industries.*

*Below we will explore how to manipulate affine transformations on the plane, which will enhance our understanding of their mathematical significance in practical applications. Using Sage to illustrate these concepts, we will focus on two-dimensional graphics (2D graphics) through the problems analyzed below. For additional learning and Sage examples, refer to the tasks outlined in 1.G.52, 1.G.53, and ??.*

**1.E.17. Affine transformations.** Find the matrix representation for simple affine transformations on the plane, as: shifts, homotheties, stretching in only one direction, shears, rotations, and reflections.

**Solution.** In order to talk about matrix representation, we need to fix the coordinates, see 1.5.2. We shall work with the standard ones, that is, with the vectors  $e_1 = (1, 0)$ ,  $e_2 = (0, 1)$ , identifying the plane with  $\mathbb{R}^2$  explicitly.

- “Identity transformation.” The simplest affine transformation is the identity, the relevant matrix is the identity matrix  $E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ . Clearly  $Eu = u$ , for all  $u \in \mathbb{R}^2$ .

- “Shifts.” The shift (or translation)  $\tau_w$  by a vector  $w = r e_1 + s e_2 = (r, s)^T \in \mathbb{R}^2$  is given by applying the matrix  $E$  and adding the vector  $w$ , i.e.,  $(x, y)^T \mapsto (x, y)^T + (r, s)^T = (x+r, y+s)^T$ , or in other words  $\tau_w(u) = u+w$ . For instance, if  $w = (3, 0)^T$ , then applying  $\tau_w$  to an image will shift it to the right by 3. Notice that when we want to shift a point  $P$  to the origin, we will also write  $\tau_{-P}$  for the relevant translation.

With the aim to illustrate a translation (and also the rest transformations), we will use a triangle  $ABC$ , with vertices the points  $A = [1, 0]$ ,  $B = [4, -2]$ , and  $C = [2, -2]$ . In the figures we have coloured these vertices by blue, green and red, respectively. The new coordinates are also painted by the corresponding colour, with the hope to get a better visualization of how the given transformation acts. For instance, let us present the illustration of the translation  $\tau_{(3,0)}$  applied to  $ABC$ :

Angles are a matter of vectors rather than points in Euclidean geometry. Let  $\mathbf{u}$  be a vector of length 1, at angle  $\varphi$  measured counter-clockwise from the vector  $(1, 0)$ . In coordinates,  $\mathbf{u}$  is at the unit circle and has first and second coordinates  $\cos \varphi, \sin \varphi$  respectively (this is one of the elementary definitions of the sine and cosine functions). That is,

$$\mathbf{u} = (\cos \varphi, \sin \varphi).$$


This is compatible with  $-1 \leq \sin \varphi \leq 1$  satisfying

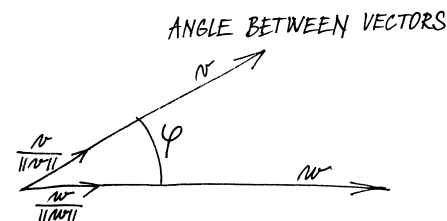
$$(\cos \varphi)^2 + (\sin \varphi)^2 = 1.$$

ANGLE BETWEEN VECTORS

The angle  $\varphi$  between two vectors  $\mathbf{v}$  and  $\mathbf{v}'$  can be in general described using their coordinates  $\mathbf{v} = (x, y)$ ,  $\mathbf{v}' = (x', y')$  like this:

$$(1) \quad \cos \varphi = \frac{xx' + yy'}{\|\mathbf{v}\| \|\mathbf{v}'\|}$$

with  $0 \leq \varphi \leq \pi$ .



In our special case  $\mathbf{v} = (1, 0)$ , the more general equation gives

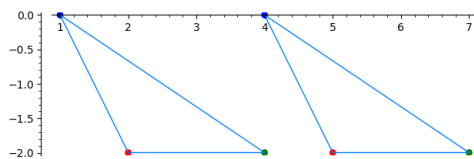
$$\cos \varphi = \frac{x'}{\|\mathbf{v}'\|},$$

which is just the definition of the function  $\cos \varphi$ . The general case can be always reduced to this special one. First we notice that the angle  $\varphi$  between two vectors  $\mathbf{u}, \mathbf{v}$  is always the same as the angle between the *normalized vectors*  $\frac{1}{\|\mathbf{u}\|} \mathbf{u}$  and  $\frac{1}{\|\mathbf{v}\|} \mathbf{v}$ . Thus we can restrict ourselves to two vectors on the unit circle. Then we can rotate our coordinates in such a way that the first of the vectors will become  $(1, 0)$ . This means, it is enough to show that the scalar product  $xx' + yy'$ , as well as the length of vectors are invariant with respect to rotations. We shall come back to this in a moment.

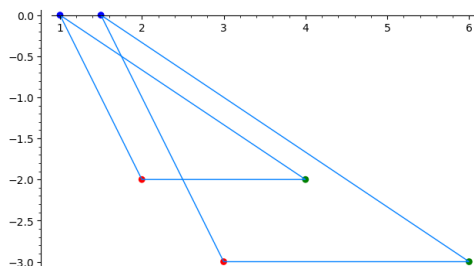
In the special case, when the scalar product is zero, we say that the vectors are *perpendicular* and this corresponds to  $\varphi = \pi/2$  as expected. Of course the best example of perpendicular vectors of length 1 are the standard basis vectors  $(1, 0)$  and  $(0, 1)$ .

Notice that our formula for the angle between the vectors is symmetric in the two vector arguments, thus we have to take the smaller of the possible angles between two vectors and  $\varphi$  is always between 0 and  $\pi$ .

We can easily imagine that not all affine coordinates are adequate for expressing the distance and thus for use in the Euclidean plane. Indeed, although we may choose any point  $O$  as the origin again, we want also the basis vectors  $\mathbf{e}_1 = \overrightarrow{OE_1}$  and  $\mathbf{e}_2 = \overrightarrow{OE_2}$  to be perpendicular and of length one. Such



• “Homotheties.” A homothety  $h_c$  is a constant rescaling of all vectors by a factor  $c \in \mathbb{R}$ . Thus the relevant matrix is a multiple of  $E$  and hence diagonal,  $A = cE = \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix}$ . As a result, a homothety with  $c > 1$  will stretch a figure. Let us illustrate this, by using the triangle  $ABC$  and the homothety  $h_c((x, y)^T) = (\frac{3x}{2}, \frac{3y}{2})^T$ , i.e.,  $c = 3/2$ :



In the opposite, a homothety  $h_c$  with  $0 < c < 1$  will shrink the figure, try to present examples yourselves.

In case we want to apply a homothety with another fixed point  $P = [a, b]$  in  $\mathbb{R}^2$  (thus rescaling the differences  $Q - P$  for all points  $Q$ ), then we can first translate  $P$  to the origin, then to employ the usual homothety and then to shift the plane back:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x - a \\ y - b \end{pmatrix} \mapsto \begin{pmatrix} c(x - a) \\ c(y - b) \end{pmatrix} \mapsto \begin{pmatrix} c(x - a) + a \\ c(y - b) + b \end{pmatrix}.$$

Imagine what is happening with the plane under the three transformations on the way.

• “Rotations.” If we rotate by angle  $\theta$  counterclockwise, we get  $e_1 \mapsto (\cos \theta, \sin \theta)$  (this is the elementary definition of the goniometric functions), while  $e_2 \mapsto (-\sin \theta, \cos \theta)$  (look at the picture in 1.5.8). Thus, the rotation matrix has the form

$$R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

For instance, for  $\theta = \pi/2$ , we get  $R_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ , and hence in this case the rotation takes  $(x, y)^T$  to  $(-y, x)^T$ :

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto R_{\pi/2} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}.$$

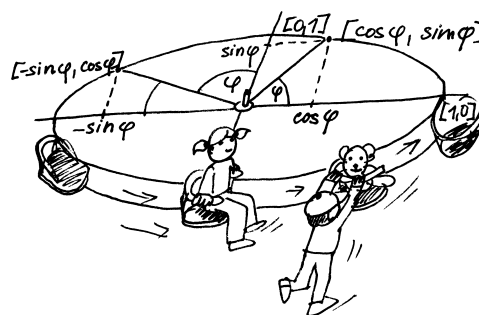
The figure below illustrates the result of  $R_{\pi/2}$  on the triangle  $ABC$  used above.<sup>10</sup>

<sup>10</sup>Notice a rotation preserves both the lengths of vectors and the angles.

basis will be called *orthonormal*. We shall see that the angles and distances computed in such coordinates will always be the same no matter which coordinates are used.

**1.5.8. Rotation around a point in the plane.** As we saw, the matrix of any given linear mapping  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is easy to guess. If the mapping works applying the matrix with columns  $(a, c)$  and  $(b, d)$ , then the first column  $(a, c)$  is obtained by multiplying this matrix with the basis vector  $(1, 0)$  and the second is the evaluation at the second basis vector  $(0, 1)$ .

ROTATION AROUND A POINT IN THE PLANE



We can see from the picture that the columns of the matrix corresponding to rotating counter-clockwise through the angle  $\psi$  are computed as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos \psi \\ \sin \psi \end{pmatrix}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin \psi \\ \cos \psi \end{pmatrix}$$

The counter-clockwise direction is called the *positive direction*, the other direction is the negative direction.

ROTATION MATRIX

Rotation through a given angle  $\psi$  in the positive direction about the origin is given by the matrix  $R_\psi$ :

$$\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} \mapsto R_\psi \cdot \mathbf{v} = \begin{pmatrix} \cos \psi & -\sin \psi \\ \sin \psi & \cos \psi \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}.$$

Now, since we now know how the matrix of the rotation in the plane looks like, we can check that rotation preserves distances and angles (defined by the equation (1) in 1.5.7).

Denote the image of a vector  $\mathbf{v}$  as

$$\mathbf{v}' = \begin{pmatrix} x' \\ y' \end{pmatrix} = R_\psi \cdot \mathbf{v} = \begin{pmatrix} x \cos \psi - y \sin \psi \\ x \sin \psi + y \cos \psi \end{pmatrix},$$

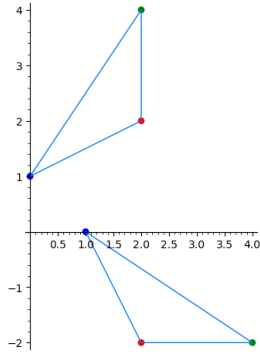
and similarly  $\mathbf{w}' = R_\psi \cdot \mathbf{w}$  for  $\mathbf{w} = (r, s)^T$ , and  $\mathbf{w}' = (r', s')^T$ .

A straightforward expansion (exploiting  $(\cos \psi)^2 + (\sin \psi)^2 = 1$ ) checks the expected equalities  $\|\mathbf{v}'\|^2 = \|\mathbf{v}\|^2$ , and that  $x'r' + y's' = xr + ys$ .

The previous expression can be written using vectors and matrices as follows:

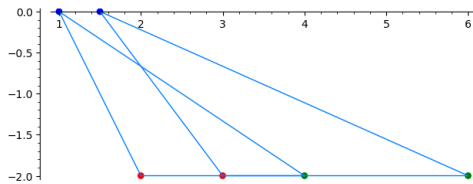
$$(R_\psi \cdot \mathbf{w})^T (R_\psi \cdot \mathbf{v}) = \mathbf{w}^T \mathbf{v}.$$



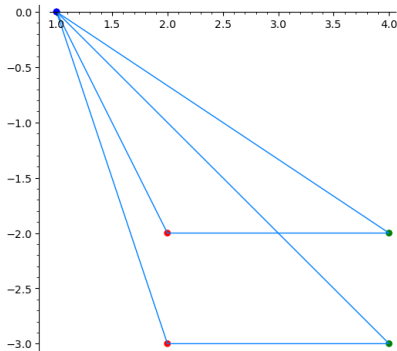


• “Stretching by a parameter  $c$ ”. Let us stretch the plane only in the direction of the first basis vector  $e_1$ . Then, the relevant matrix  $A_c$  is diagonal, obtained by multiplying the first row in the identity matrix by the stretching parameter  $c$ , i.e.,  $A_c = \begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix}$ . The figure below shows the stretching of our triangle  $ABC$  in the direction of  $e_1$  with  $c = 3/2$ , hence the particular linear transformation has the form

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 3/2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3x/2 \\ y \end{pmatrix}.$$



Applying the same stretching but in direction of  $e_2$ , will force only the  $y$ -coordinates to be transformed, as we see here:



If we want to employ the same operation with a general fixed point  $P$  belonging in a line  $\ell$ , and along this line, we may again compose three transformations: first shift by  $\tau_{-P}$ , then find the angle  $\theta$  of  $\ell$  with the direction of  $e_1$  and use the  $R_{-\theta}$  matrix, then apply the above stretching and next move the plane back, i.e., apply  $R_\theta$  and  $\tau_P$ .

• “Shears.” The horizontal shear is a linear transformation keeping the  $y$ -coordinate fixed, and the entire “ $x$ -coordinate axes” (this is the line  $y = 0$ ), is fixed, too. In other words a horizontal shear takes the point  $(x, y)$  to the point  $(x + ay, y)$  for some  $a$ . Similarly, vertical shears are the transformations with the properties of the components swapped. Thus, the relevant matrices have the following

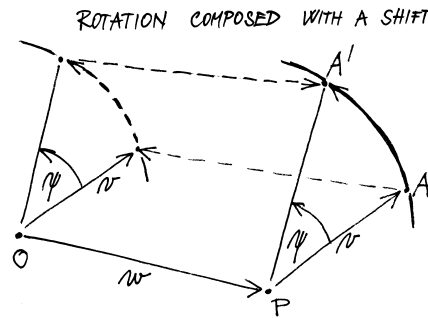
The transposed vector  $(R_\psi \cdot \mathbf{w})^T$  equals  $\mathbf{w}^T \cdot R_\psi^T$ , where  $R_\psi^T$  is the so-called transpose of the matrix  $R_\psi$ . That is a matrix, whose rows consist of the columns of the original matrix and similarly the columns consist of the rows of the original matrix. Therefore we see that the rotation matrices satisfy the relation  $R_\psi^T \cdot R_\psi = I$ , where the matrix  $I$  (sometimes we denote this matrix just as 1 and mean by this the unit in the ring of matrices) is the *unit matrix*

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This leads us to a remarkable observation — the matrix  $X$  with the property that  $X \cdot R_\psi = I$  (we will call such a matrix the *inverse matrix to the rotation matrix*  $R_\psi$ ) is the transpose of the original matrix. This makes sense, since the inverse mapping to the rotation through the angle  $\psi$  is again a rotation, but through the angle  $-\psi$ . That is, the inverse matrix of  $R_\psi^T$  equals the matrix

$$R_{-\psi} = \begin{pmatrix} \cos(-\psi) & -\sin(-\psi) \\ \sin(-\psi) & \cos(-\psi) \end{pmatrix} = \begin{pmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{pmatrix}.$$

It is easy to write the rotation around a point  $P = O + \mathbf{w}$ ,  $P = [r, s]$  again using matrices. One just has to note that instead of rotating around the given point  $P$ , we can first shift  $P$  into the origin, then do the rotation and then do the inverse shift.



We calculate:

$$\begin{aligned} \mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix} &\mapsto \mathbf{v} - \mathbf{w} \mapsto R_\psi \cdot (\mathbf{v} - \mathbf{w}) \\ &\mapsto R_\psi \cdot (\mathbf{v} - \mathbf{w}) + \mathbf{w} = \\ &\begin{pmatrix} \cos \psi (x - r) - \sin \psi (y - s) + r \\ \sin \psi (x - r) + \cos \psi (y - s) + s \end{pmatrix}. \end{aligned}$$

**1.5.9. Reflections.** Another well-known example of a length preserving mapping is *reflection through a line*. It is enough to understand reflection through a line that goes through the origin  $O$ . All other reflections can be derived using shifts. Thus, we look first for a matrix  $Z_\psi$  of reflection with respect to the line through the origin and through the point  $(\cos \psi, \sin \psi)$ .



Notice that

$$Z_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

form:

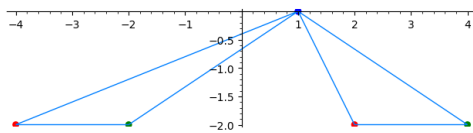
$$H_a = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, \quad V_a = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}.$$

Let us illustrate a shear in the  $x$ -direction for the values  $a = \pm 3$ . We get the transformations

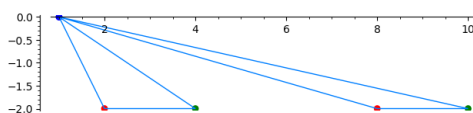
$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 1 & \pm 3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \pm 3y \\ y \end{pmatrix},$$

so the triangle  $ABC$  is transformed as follows:

Case  $a = 3$ :

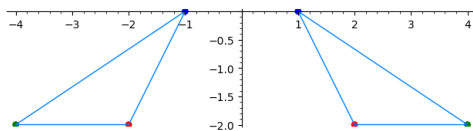


Case  $a = -3$ :



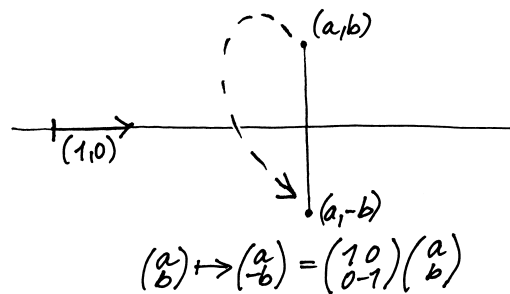
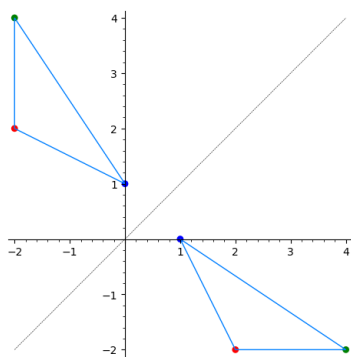
- “Reflections.” Finally, let us deal with reflections.

They actually appeared above as stretching along one direction, with the parameter  $c = -1$ . Thus, stretching along the direction of  $e_1$  with  $c = -1$  gives a reflection with respect to the  $y$ -axis, where the matrix representation is  $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ . Repeating, using instead of  $e_1$  the vector  $e_2$ , we get a reflection with respect to the  $x$ -axis, where the matrix presentation is  $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ . Below we illustrate the reflection of our triangle  $ABC$  with respect to the  $y$ -axis:



For an illustration of a reflection in the plane with respect to  $x$ -axis, see 1.G.51.

On the other hand, the reflection in the plane through the line  $y = x$ , sends the  $x$ -axis into the  $y$ -axis, and vice versa. Hence it is also referred to as the “axial symmetry”. Obviously, its matrix presentation is given by  $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ , hence for example it transforms the given triangle as follows (see 1.G.52 for another example).



Any line going through the origin can be rotated so that it has the direction  $(1, 0)$  and thus we can write general reflection matrix as

$$Z_\psi = R_\psi \cdot Z_0 \cdot R_{-\psi},$$

where we first rotate via the matrix  $R_{-\psi}$  so that the line is in “zero” position, reflect with the matrix  $Z_0$  and return back with the rotation  $R_\psi$ .

Therefore we can calculate (by associativity of matrix multiplication):

$$\begin{aligned} Z_\psi &= \begin{pmatrix} \cos \psi & -\sin \psi \\ \sin \psi & \cos \psi \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{pmatrix} \\ &= \begin{pmatrix} \cos \psi & \sin \psi \\ \sin \psi & -\cos \psi \end{pmatrix} \cdot \begin{pmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{pmatrix} \\ &= \begin{pmatrix} \cos^2 \psi - \sin^2 \psi & 2 \sin \psi \cos \psi \\ 2 \sin \psi \cos \psi & -(\cos^2 \psi - \sin^2 \psi) \end{pmatrix} \\ &= \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ \sin 2\psi & -\cos 2\psi \end{pmatrix}. \end{aligned}$$

The last equality follows from the usual formulas for trigonometric functions:

$$(1) \quad \begin{aligned} \sin 2\psi &= 2 \sin \psi \cos \psi \\ \cos 2\psi &= \cos^2 \psi - \sin^2 \psi. \end{aligned}$$

Notice that the product  $Z_\psi \cdot Z_0$  gives:

$$\begin{pmatrix} \cos 2\psi & \sin 2\psi \\ \sin 2\psi & -\cos 2\psi \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \cos 2\psi & -\sin 2\psi \\ \sin 2\psi & \cos 2\psi \end{pmatrix}.$$

This observation can be formulated as follows:

**Proposition.** A rotation through the angle  $\psi$  can be obtained by two subsequent reflections through the lines that have the angle  $\frac{1}{2}\psi$  between them.  $\square$

**1.E.18.** Show that every rotation can be obtained as a composition of three shears.<sup>11</sup>

**Solution.** This is a little more demanding exercise.



But we may experiment with the abilities of Sage and this makes our life really easy. We would add the lines in the following session following partial results. Composing horizontal shears gives again a horizontal shear (and similarly for the vertical ones):

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & a+b \\ 0 & 1 \end{pmatrix}.$$

Thus we try to combine horizontal and vertical ones to get a rotation by an angle  $\theta$ .

```
# introduce three shear matrices with
# variable parameters a,b,c
a,b,c=var("a,b,c")
A = matrix([[1,a],[0,1]]); B = matrix(
[[1,0],[b,1]]); C=matrix([[1,c],[0,1]])
# take the general transformation X
# obtained by composition
X = C*B*A; print(X)
# notice that one of the conditions on X
# to be a rotation indicates a=c
X = X.substitute(c=a)
eq1 = X[0,0]^2+X[1,0]^2-1;
eq2 = X[1,0]+X[0,1];
# eq3 = X[0,0]-X[1,1] == 0 is already
# fulfilled; solve the system
sol = solve([eq1,eq2],[a,b]); print(sol)
```

The final line returns the two solutions:

$$a = (\sqrt{-r_1^2 + 1} - 1)/r_1, \quad b = r_1,$$

or

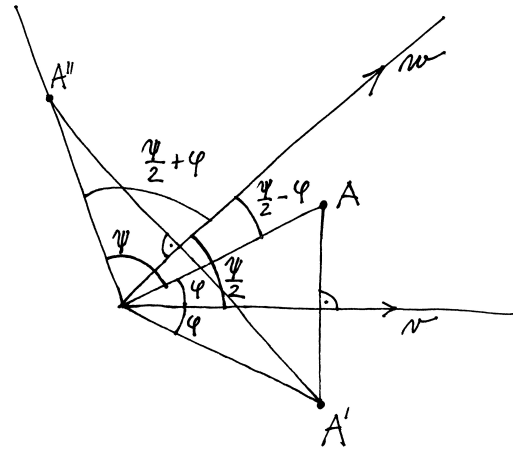
$$a = -(\sqrt{-r_2^2 + 1} + 1)/r_2, \quad b = r_2.$$

This means that  $b$  is a free parameter and from the printed form of  $X$ , we learn that  $b$  plays the role of  $\sin \theta$ . The graphical illustration of this compositions is given in ??.

**1.E.19. Change of coordinates.** Look at picture in 1.5.6 and assume that, in the standard coordinates,  $O' = [1, 5]$ ,  $P = O' + e'_1 = [2, 3]$ ,  $Q = O' + e'_2 = [2, 7]$  and these define the “girl’s” coordinate system. Consider the two options that the meeting point  $M = [2, 1]$  is given (a) by the boy, or (b) by the girl, in their own coordinate system. What does it mean for the other one? Find the answer first for the situation where at least the origins of the coordinate systems coincide!

**Solution.** If the boy and girl have got at least the same origin of coordinate systems, then the transformation of coordinates is simple. The girl’s basis is already expressed in the standard coordinates (the boy’s ones). Thus the matrix  $T$  with columns

<sup>11</sup>This is actually of high interest for pixel based computer graphics: A horizontal or vertical shear is a very quick operation consisting of constant shifting of individual rows or columns of pixels, while implementing a rotation directly is much more time consuming! It seems that the first algorithm based on this was introduced by Alan Paeth in 1986.



In fact we can prove the previous proposition purely by geometrical argumentation, as shown in the above picture (try to be a “synthetic geometer” when reflecting  $A$  to  $A'$  and then  $A'$  to  $A''$ ). If we believe in this proof “by picture”, then the above computational derivation of the proposition provides the proof of the standard double angle formulas (1).

**1.5.10.** The following is a recapitulation of previous ideas.

MAPPINGS THAT PRESERVE LENGTH

**Theorem.** A linear mapping of the Euclidean plane is composed of one or more reflections if and only if it is given by a matrix  $R$  which satisfies

$$R = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad ab + cd = 0, \quad a^2 + c^2 = b^2 + d^2 = 1.$$

This happens if and only if the mapping preserves length. Rotation is such a mapping if and only if the determinant of the matrix  $R$  equals one, which corresponds to an even number of reflections. When there is an odd number of reflections, the determinant equals  $-1$ .

**PROOF.** We calculate how a general matrix  $A$  might look, when the corresponding mapping preserves length. That is, we have a mapping



$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}.$$

Preserving length thus means that for every  $x$  and  $y$ , we have

$$\begin{aligned} x^2 + y^2 &= (ax + by)^2 + (cx + dy)^2 \\ &= (a^2 + c^2)x^2 + (b^2 + d^2)y^2 + 2(ab + cd)xy. \end{aligned}$$

Since this equation is to hold for every  $x$  and  $y$ , the coefficients of the individual powers  $x^2$ ,  $y^2$  and  $xy$  on the left and right side of the equation must be equal. Thus we have calculated that the conditions put on the matrix  $R$  in the first part

$e'_1$  and  $e'_2$  yields the linear mapping providing the boy's coordinate expression of the girl's vectors. In order to go in the opposite direction, we just take the inverse matrix  $T^{-1}$ . Thus,

$$T = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}, \quad \text{and} \quad T^{-1} = \begin{pmatrix} 1/2 & -1/4 \\ 1/2 & 1/4 \end{pmatrix},$$

and the boy knows that the girl means him to be at  $[3, -2]$ , while the girl knows that boy's  $[2, 1]$  would be  $[3/4, 5/4]$ .

In the general case, we must consider the shifts of the origin, evaluated in the right coordinates. We shall again use Sage to compute:

```
# go to the new affine coordinates
O=vector([0,0]);e1=vector([1,0]);
e2=vector([0,1]);
OO=vector([1,5]);ee1=vector([2,3])-OO;
ee2 = vector([2,7])-OO;
# define the relevant transformations
T=matrix([ee1,ee2]); T=T.transpose();
print(T); print(T.inverse())
def change_girl_to_boy(coordinates):
    position = vector(coordinates)
    return T*position + OO
def change_boy_to_girl(coordinates):
# [0,0] = a[1,-2]+b[1,2]+[1,5] =>
# a-b=5/2, a+b=-1 => a=3/4, b=-7/4
    position = vector(coordinates)
    return T.inverse()*position
    + vector([3/4,-7/4])
print(change_girl_to_boy(vector([2,1])))
print(change_boy_to_girl(vector([2,1])))
# tests
OO+2*ee1+1*ee2 == O+4*e1+3*e2 and
O+2*e1+1*e2 == OO+3/2*ee1-1/2*ee2
```

The linear parts of the transformations are still  $T$  and  $T^{-1}$ , the right shifts are added. The last two prints give the answer: the girl's meeting point  $[2, 1]$  means the boy should go to position  $[4, 3]$ , while if it is the boy who tells his meeting point, then the girl should look for him at  $[3/2, -1/2]$  in her coordinates. The final line confirms this by returning "True".  $\square$

*Let us now move to Euclidean tasks. In addition, there is the concept of the norm (size) of a vector, allowing us to measure distances between points. We have already seen this in case of complex numbers, viewing  $\mathbb{C}$  as the real plane with standard coordinates. As we will see below, we are then also able to measure angles between vectors and areas of polygonal objects.*



- 1.E.20.** (a) Find the norms, the scalar product and the angle of the vectors  $u = (-3, -2)$  and  $v = (-2, 3)$ .  
 (b) Prove that the triangle with vertices  $P = [2, 2], Q = [3, 0], R = [4, 3]$  is isosceles.

**Solution.** (a) We shall work with the standard Euclidean norm on  $\mathbb{R}^2$  (which we may identify with  $\mathbb{C}$ ), as we did in 1.A.8. The norm of a vector  $w = (x, y)$  is given by

of the theorem we are proving are equivalent to the property that the given mapping preserves length.

Because  $a^2 + c^2 = 1$ , we can assume that  $a = \cos \varphi$  and  $c = \sin \varphi$  for a suitable angle  $\varphi$ . As soon as we choose the first column of the matrix  $R$ , the relation  $ab + cd = 0$  determines the second column up to a multiple. But we also know that the length of the vector in the second column is one, and thus we have only two possibilities for the matrix  $R$ , namely:

$$\begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}, \quad \begin{pmatrix} \cos \varphi & \sin \varphi \\ \sin \varphi & -\cos \varphi \end{pmatrix}.$$

In the first case, we have a rotation through the angle  $\varphi$ , in the second case we have a rotation composed with the reflection through the first coordinate axis. As we have seen in the previous proposition 1.5.8, every rotation corresponds to two reflections. The determinant of the matrix  $R$  is in these two cases either one or minus one and distinguishes between these two cases by the parity of the number of reflections.  $\square$

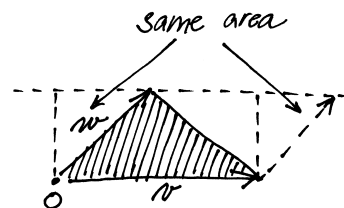
Notice, we have now proved our earlier claim on the invariance of formulae for distance and angle in any orthonormal coordinates. Moreover, we have seen that all euclidean affine mappings are generated by translations, and reflections.

**1.5.11. Area of a triangle.** At the end of our little trip to



geometry we will focus on the *area* of planar objects. For us, triangles will be sufficient. Every triangle is determined by a pair of vectors  $\mathbf{v}$  and  $\mathbf{w}$ , which, if translated so that they start from one vertex  $P$  of the triangle, determine the remaining two vertices. We would like to find a formula (scalar function area), which assigns the number  $\text{area } \Delta(\mathbf{v}, \mathbf{w})$  equal to the area of the triangle  $\Delta(\mathbf{v}, \mathbf{w})$  defined in the aforementioned way. By translating, we can place  $P$  at the origin since translation should not change the area.

*area of  $\Delta = 1/2$  area of  $\square$*



We can see from the picture that the desired value is half of the area of the parallelogram spanned by the vectors  $\mathbf{v}$  and  $\mathbf{w}$ . It is easy to calculate (using the well-known formula: base times corresponding height), or simply observe from the picture that the following holds

$$\begin{aligned} \text{area } \Delta(\mathbf{v} + \mathbf{v}', \mathbf{w}) &= \text{area } \Delta(\mathbf{v}, \mathbf{w}) + \text{area } \Delta(\mathbf{v}', \mathbf{w}) \\ \text{area } \Delta(a\mathbf{v}, \mathbf{w}) &= a \text{ area } \Delta(\mathbf{v}, \mathbf{w}). \end{aligned}$$

$\|w\| = \sqrt{x^2 + y^2}$ . Thus  $\|u\| = \sqrt{9+4} = \sqrt{13} = \|v\|$ . Their scalar product (dot product) is given by  $u \cdot v \equiv \langle u, v \rangle = (-3)(-2) + (-2)3 = 0$ , and hence  $u, v$  are orthogonal each other, i.e., their angle is  $\varphi = \pi/2$ . In Sage, we have got the method which is applied to the first vector and gets the second one as argument: `u.dot_product(v)`, see the cell

```
u = vector([-3, 2]); v = vector([-2, 3])
u.dot_product(v)
```

which returns 0. As for the norms of the vectors  $u, v$ , type

```
u.norm(); v.norm()
```

and for both cases this returns  $\sqrt{13}$ .

(b) In this case we compute

$$\begin{aligned} |\overrightarrow{PQ}| &= \|P - Q\| = \sqrt{(2-3)^2 + (2-0)^2} = \sqrt{5}, \\ |\overrightarrow{QR}| &= \|Q - R\| = \sqrt{(3-4)^2 + (0-3)^2} = \sqrt{10}, \\ |\overrightarrow{PR}| &= \|P - R\| = \sqrt{(2-4)^2 + (2-3)^2} = \sqrt{5}. \end{aligned}$$

Thus  $|\overrightarrow{PQ}| = |\overrightarrow{PR}|$  and the triangle is isosceles. □

**1.E.21.** Given the line  $\ell : 7y = 6x + 13$ , determine the line  $\ell'$  which is perpendicular to  $\ell$  and passes through the point  $P = [-6, 7] \in \mathbb{R}^2$ . Find at least two methods.

**Solution.** (1) Given a line  $\ell$  in parametric form  $P_0 + tu = [x_0, y_0] + t(x_1, y_1)$  ( $t \in \mathbb{R}$ ), it is easy to determine a line  $\ell'$  perpendicular to  $\ell$ , which goes through another point  $Q = [a, b]$ , as follows. The direction vector of  $\ell'$  must be orthogonal with the direction vector of  $\ell$ . So, if we assume that  $Q + tv$  is the parametric form of  $\ell'$ , then we should have  $u \cdot v = 0$ , where  $\cdot$  is the usual dot product on  $\mathbb{R}^2$ . Taking  $v = (-y_1, x_1)$  we see that  $(x_1, y_1) \cdot (-y_1, x_1) = 0$ , that is,  $u \perp v$ . Thus the parametric equation of  $\ell'$  is simply  $Q + tv = [a, b] + t(-y_1, x_1)$ , with  $t \in \mathbb{R}$ . To apply this in our case we need to express the initial line  $\ell$  in parametric form. The points  $[-1, 1]$  and  $[0, 13/7]$  belong to the line  $\ell$ , so the parametric form of  $\ell$  is (cf. 1.E.8)

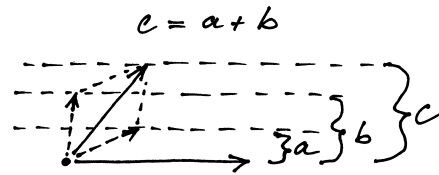
$$[-1, 1] + t(0 - (-1), \frac{13}{7} - 1) = [-1, 1] + t(1, \frac{6}{7}).$$

As a result, the direction vector of  $\ell'$  is  $(-\frac{6}{7}, 1)$ , such that  $(1, \frac{6}{7}) \cdot (-\frac{6}{7}, 1) = 0$ , and  $\ell'$  reads by  $[-6, 7] + t(-\frac{6}{7}, 1)$ , or in other words  $x = -6 - \frac{6}{7}t$  and  $y = 7 + t$ , with  $t \in \mathbb{R}$ . For a verification we can transfer the given parametric form of  $\ell'$  to implicit form. Eliminating  $t$  from these equations gives  $x = -6 - \frac{6}{7}(y - 7)$ , which is equivalent to  $y = -\frac{7}{6}x$ .

(2) Actually, writing the equation for  $\ell$  as  $6x - 7y - 13 = 0$ , it is obvious that the coefficients  $(6, -7)$  provide just a vector perpendicular to  $\ell$ . Thus we may write the implicit equation for  $\ell'$  immediately.

(3) We may also work with the slopes. Writing  $\ell$  as  $y = \frac{6}{7}x + \frac{13}{7}$ , we get its slope  $m = 6/7$ . The line  $\ell'$ , perpendicular to the line  $\ell$  must have the slope  $m' = -\frac{7}{6}$ , such that  $mm' = -1$ .

LINEARITY IN ARGUMENT



Finally we add to the formulation of our problem a condition

$$\text{area } \Delta(\mathbf{v}, \mathbf{w}) = -\text{area } \Delta(\mathbf{w}, \mathbf{v}),$$

which corresponds to the idea that we give a sign to the area, according to the order in which we are taking the vectors.

If we write the vectors  $\mathbf{v}$  and  $\mathbf{w}$  into the columns of a matrix  $A$ , then the mapping

$$A = (\mathbf{v}, \mathbf{w}) \mapsto \det A$$

satisfies all the three conditions we wanted. How many such mappings could there possibly be? Every vector can be expressed using two basis vectors  $\mathbf{e}_1 = (1, 0)$  and  $\mathbf{e}_2 = (0, 1)$ . By linearity, area  $\Delta$  is uniquely determined by these vectors. We want

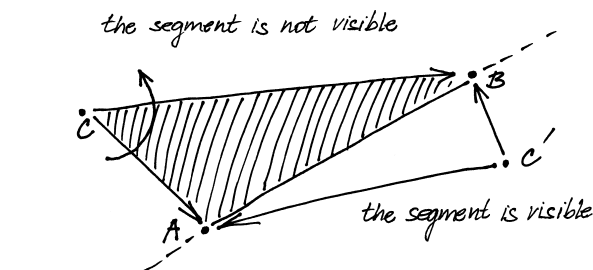
$$\text{area } \Delta(\mathbf{e}_1, \mathbf{e}_2) = \frac{1}{2}.$$

In other words, we have chosen the *orientation* and the *scale* through the choice of basis vectors, and we choose the unit square to have area equal to one.

Thus we see that the determinant gives the area of a parallelogram determined by the columns of the matrix  $A$ . The area of the triangle is thus one half of the parallelogram.

**1.5.12. Visibility in the plane.** The previous description of the value for oriented area gives us an elegant tool for determining the position of a point relative to oriented line segments. By an oriented line segment we mean two points in the plane  $\mathbb{R}^2$  with a selected order. We can imagine it as an arrow from one point to the other. Such an oriented line segment divides the plane into two half-planes. Let us call them “left” and “right”. We want to be able to determine whether a given point is in the left or right half-plane.

Such tasks are often met in computer graphics when dealing with visibility of objects. We can imagine that an oriented line segment can be “seen” from the points to the right of it and cannot be seen from the points to left of it.



Thus,  $\ell'$  is of the form  $y = -\frac{7}{6}x + b$ , and substituting the coordinates of  $P$  implies  $b = 0$ , i.e.,  $y = -\frac{7}{6}x$ .  $\square$

**1.E.22.** Verify that each affine mapping always maps parallel lines to parallel lines, but the angles and distances are changed in general.  $\circ$

**1.E.23. Remark on plots.** Sage, as well as other software, provides extensive 2D plotting functionality. However, even for drawing simple lines in the plane, we should notice several details.

First, let us look at the so called “aspect ratio” of the figures, which is governed by the option `aspect_ratio` in Sage. The aspect ratio reflects the apparent height/width ratio of a unit square. In Sage one can explicitly ask for an “automatic setup” by typing `aspect_ratio = "automatic"`. However, this does not produce always what we normally expect. Instead, the option `aspect_ratio = 1` works much better applying the same scale on both coordinates. This seems to be relevant if illustrating Euclidean properties, where we want to see the real angles and lengths, while in the affine setup we may more appreciate to choose the scales independently.

We may also adjust the size of a figure via the option `figsize`. The default `figsize` is 4, while setting `figsize = 8` makes a figure roughly twice as big. We can also request separate horizontal and vertical sizes, by typing for example `figsize = [4, 8]` (both will be measured in inches). Usually, small `figsize` values work well.

**1.E.24.** To illustrate the above situation, provide the plots of the lines appearing in Problem 1.E.21.  $\circ$

*In the next series of exercises we shall focus on applications of the linear and affine mappings. In applications, they appear either directly, or as “linear approximations”.<sup>12</sup> We already saw the simple building blocks of most of them in 1.E.17, see also the other column starting in 1.5.8.*

*We shall also include the concept of area and its link to the properties of the determinant function on matrices. We refer to the paragraphs 1.5.5 and 1.5.11 for details.*

**1.E.25.** Compute the area of the triangle  $PQR$ , if  $P = [-8, 1]$ ,  $Q = [-2, 0]$  and  $R = [5, 9]$ . What would we have to do, if the points were given in a coordinate system different to the standard Euclidean one?  $\circ$

**1.E.26.** Consider the quadrilateral  $S$  given by its vertices  $A = [1, 1]$ ,  $B = [6, 1]$ ,  $C = [11, 4]$ , and  $D = [2, 4]$ . Illustrate  $S$  via Sage and deduce that  $S$  is a trapezoid whose bases are of lengths 5 and 9, respectively, and whose height equals to 3. Next compute its area in at least two ways.  $\circ$

<sup>12</sup>While the linear approximations are the core of building models in Science and Engineering (we shall start developing the relevant methods in Chapter 5), in arts, the linearity concept is essential in finding the right proportions, helping to create structure, balance, and rhythm in compositions, guiding the viewer’s eye and effecting emotions.

We have the oriented line segment  $AB$  and are given some point  $C$ . We calculate the oriented area of the corresponding triangle determined by the vectors  $A-C$  and  $B-C$ . If the point  $C$  is to the left of the line segment, then with the usual positive orientation (counter-clockwise) we obtain the positive sign of the oriented area, while the negative sign corresponds to the points to the right.

This approach is often used for testing relative positions in 2D graphics.

## 6. Relations and mappings

In the final part of this introductory chapter, we return to the formal description of mathematical structures. We will try to illustrate them on examples we already know. We can consider this part to be an exercise in a formal approach to the objects and concepts of mathematics, i.e. the “language of mathematics”.



**1.6.1. Relations between sets.** First we define the *Cartesian product*  $A \times B$  of two sets  $A$  and  $B$ . It is the set of all ordered pairs  $(a, b)$  such that  $a \in A$  and  $b \in B$ . A *binary relation* between the two sets  $A$  and  $B$  is then a subset  $R$  of the Cartesian product  $A \times B$ .

We write  $a \simeq_R b$  to mean  $(a, b) \in R$ , and say that  $a$  is related to  $b$ . The *domain of the relation* is the subset

$$D = \{a \in A : \exists b \in B, (a, b) \in R\}.$$

Here the symbol  $\exists b$  means that there is at least one such  $b$  satisfying the rest of the claim.

Similarly, the *codomain of the relation* is the subset

$$I = \{b \in B : \exists a \in A, (a, b) \in R\}.$$

A special case of a relation between sets is a *mapping from the set  $A$  to the set  $B$* . This is the case when every element of the domain of the relation is related to exactly one element of the codomain. Examples of mappings known to us are all functions, where the codomain of the mapping is a set of numbers, for instance the set of integers or the set of real numbers, or the linear mappings in the plane given by matrices. We write



$$f : D \subseteq A \rightarrow I \subseteq B, \\ f(a) = b$$

to express the fact that  $(a, b)$  belongs to a relation, and we say that  $b$  is the value of  $f$  at  $a$ . Furthermore we say that

- mapping  $f$  of the set  $A$  to the set  $B$  is *surjective* (or *onto*), if  $D = A$  and  $I = B$ , clarify ?
- mapping  $f$  of the set  $A$  to the set  $B$  is *injective* (or *one-to-one*), if  $D = A$  and for every  $b \in I$  there exist exactly one *preimage*  $a \in A$ ,  $f(a) = b$ .

Expressing a mapping  $f : A \rightarrow B$  as a relation

$$f \subseteq A \times B, \quad f = \{(a, f(a)); a \in A\}$$

is also known as *the graph of a mapping  $f$* .

**1.E.27.** (a) An equilateral triangle with vertices  $[1, 0]$  and  $[0, 1]$  lies entirely in the first quadrant. Find the coordinates of its third vertex with the help of Sage.

(b) An equilateral triangle has vertices at  $P = [1, 1]$  and  $Q = [2, 3]$ . Its other vertex lies in the same half-plane as the point  $S = [0, 0]$ . The triangle is rotated by  $60^\circ$  in the positive direction around the point  $S$ , to produce a new triangle. Determine the coordinates of the vertices of the new triangle.

**Solution.** (a) Let us denote by  $A = [1, 0]$  and  $B = [0, 1]$  the given points. To find the coordinates  $[x, y]$  of the third vertex we are rotating the point  $A$  around the point  $B$  by an angle of  $60^\circ = \pi/3$  in the positive direction (counter clockwise direction). Thus, in Sage we can find the third vertex by typing

```
a=pi/3
M=matrix([[cos(a), -sin(a)], [sin(a), cos(a)]])
A=vector([1, 0]); B=vector([0, 1])
rot=M*(A-B)+B;
print(rot)
```

Hence we have

$$[x, y] = \begin{pmatrix} \cos(\pi/3) & -\sin(\pi/3) \\ \sin(\pi/3) & \cos(\pi/3) \end{pmatrix} (A - B) + B,$$

which gives  $[x, y] = [\frac{1}{2} + \frac{\sqrt{3}}{2}, \frac{1}{2} + \frac{\sqrt{3}}{2}]$ .

(b) Using the general rotation transformation as found in Sage above, we obtain easily the wanted points (check yourselves):

$$[-\frac{3}{2}\sqrt{3}, \sqrt{3} - \frac{1}{2}], [\frac{1}{2} - \frac{1}{2}\sqrt{3}, \frac{1}{2}\sqrt{3} + \frac{1}{2}], [1 - \frac{3}{2}\sqrt{3}, \sqrt{3} + \frac{3}{2}].$$

□

**1.E.28.** Consider a regular hexagon with vertices labeled in the positive direction, with first vertex  $A = [0, 2]$  and centre at the point  $S = [1, 0]$ . Determine the coordinates of the third vertex  $C$ . ○

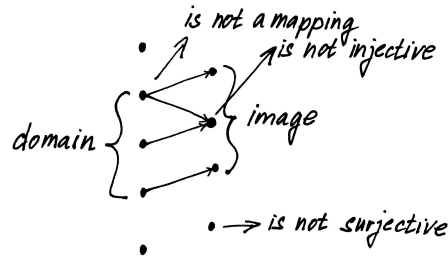
**1.E.29.** Which sides of the quadrangle given by the vertices  $[-2, -2]$ ,  $[1, 4]$ ,  $[3, 3]$  and  $[2, 1]$  are “visible” from the position of the point  $X = [3, \pi - 2]$ ?

**Solution.** In the first step we order the vertices such that their order corresponds the counter-clockwise direction. We choose vertex  $A = [-2, -2]$ , the order of the remaining vertices is then  $B = [2, 1]$ ,  $C = [3, 3]$ ,  $D = [1, 4]$  (think, how to order the points without a picture; you can actually use similar procedure to what follows). First consider the side  $AB$ . It along with the point  $X = [3, \pi - 2]$  determines the matrix

$$\begin{pmatrix} -2 - 3 & 2 - 3 \\ -2 - (\pi - 2) & 1 - (\pi - 2) \end{pmatrix}$$

such that its first column is the difference  $A - X$  and the second column is  $B - X$ . Whether it can be “seen” from the point  $[3, \pi - 2]$  (i.e., is left or right of the vector  $\overrightarrow{AB}$ , see 1.5.12), is then determined by the sign of the determinant

$$\begin{vmatrix} -2 - 3 & 2 - 3 \\ -2 - (\pi - 2) & 1 - (\pi - 2) \end{vmatrix} = \begin{vmatrix} -5 & -1 \\ -\pi & 3 - \pi \end{vmatrix} < 0.$$



**1.6.2. Composition of relations and mappings.** For mappings, the concept of composition is clear. Suppose we have two mappings  $f : A \rightarrow B$  and  $g : B \rightarrow C$ . Then their composition  $g \circ f : A \rightarrow C$  is defined as

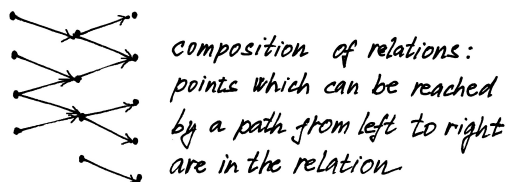
$$(g \circ f)(a) = g(f(a)).$$

Composition can also be expressed with the notation used for a relation as

$$\begin{aligned} f &\subseteq A \times B, & f &= \{(a, f(a)); a \in A\} \\ g &\subseteq B \times C, & g &= \{(b, g(b)); b \in B\} \\ g \circ f &\subseteq A \times C, & g \circ f &= \{(a, g(f(a))); a \in A\}. \end{aligned}$$

The composition of a relation is defined in a very similar way. We just add existential quantifiers to the statements, since we have to consider all possible “preimages” and all possible “images”. Let  $R \subseteq A \times B$ ,  $S \subseteq B \times C$  be relations. Then  $S \circ R \subseteq A \times C$ ,

$$S \circ R = \{(a, c); \exists b \in B, (a, b) \in R, (b, c) \in S\}.$$



A special case of a relation is the identity relation

$$\text{id}_A = \{(a, a) \in A \times A; a \in A\}$$

on the set  $A$ . It is a neutral element with respect to composition with any relation that has  $A$  as its codomain or domain.

For every relation  $R \subseteq A \times B$ , we define the inverse relation

$$R^{-1} = \{(b, a); (a, b) \in R\} \subseteq B \times A.$$

Beware, the same term is used with mappings in a more specific situation. Of course, for every mapping there is its inverse relation, but this relation is in general not a mapping. Therefore we speak about the existence of an inverse mapping if every element  $b \in B$  is an image of exactly one element in  $A$ . In such a case the inverse mapping is exactly the inverse relation.

Note that the composition of a mapping and its inverse mapping (if it exists) is the identity mapping. In general, this is not so for relations.

For the sides  $BC$ ,  $CD$  and  $DA$  we analogically obtain

$$\begin{aligned} \begin{vmatrix} 2-3 & 3-3 \\ 1-(\pi-2) & 3-(\pi-2) \end{vmatrix} &< 0, \\ \begin{vmatrix} 3-3 & 1-3 \\ 3-(\pi-2) & 4-(\pi-2) \end{vmatrix} &> 0, \\ \begin{vmatrix} 1-3 & -2-3 \\ 4-(\pi-2) & -2-(\pi-2) \end{vmatrix} &> 0. \end{aligned}$$

The determinants differ in signs, thus the point  $X$  is outside the given quadrangle and a side is visible from  $X$ , if  $X$  is left of the side. According to our convention of putting vectors  $\overrightarrow{XA}$ ,  $\overrightarrow{XB}$ ,  $\overrightarrow{XC}$ ,  $\overrightarrow{XD}$  into the determinants, the side is visible, if the corresponding determinant is negative (i.e.,  $X$  is right of the oriented side). Thus, from the point  $X$  are visible exactly the sides determined by the pairs of vertices  $A = [-2, -2]$ ,  $B = [2, 1]$  and  $B = [2, 1]$ ,  $C = [3, 3]$ .  $\square$

**1.E.30.** Given the sides of the pentagon with vertices at points  $[-2, -2]$ ,  $[-2, 2]$ ,  $[1, 4]$ ,  $[3, 1]$  and  $[2, -11/6]$ , which are visible from the point  $[300, 1]$ .  $\circ$

**1.E.31.** Let the triangle with the vertices  $P = [5, 6]$ ,  $Q = [7, 8]$ , and  $R = [5, 8]$ . Determine, which of its sides are visible from the point  $X = [0, 1]$ .  $\circ$

**F. Relations and mappings**

*We close this chapter by considering briefly some aspects of the language of mathematics. Thus, instead of the usual intuitive introduction we request the reader to have a quick look at the formal approach to the relevant concepts, beginning in 1.6.1. Note that digesting the abstract concepts of equivalence and ordering (see 1.6.3) is one of the crucial steps towards mathematical thinking.*



**1.F.1.** Determine whether the following relations on the given set  $M$  are equivalence relations:

- i)  $M = \{f : \mathbb{R} \rightarrow \mathbb{R}\}$ , where  $f \sim g$  if  $f(0) = g(0)$ .
- ii)  $M = \{f : \mathbb{R} \rightarrow \mathbb{R}\}$ , where  $f \sim g$  if  $f(0) = g(1)$ .
- iii)  $M$  is the set of lines in the plane, where two lines are related if they do not intersect.
- iv)  $M$  is the set of lines in the plane, where two lines are related if they are parallel.
- v)  $M = \mathbb{N}$ , where  $m \sim n$  if  $S(m) + S(n) = 20$ , while  $S(n)$  stands for the sum of the digits of the integer  $n$ .
- vi)  $M = \mathbb{N}$ , where  $m \sim n$  if  $C(m) = C(n)$ , where  $C(n) = S(n)$  if the sum of the digits  $S(n)$  is less than 10, otherwise we define  $C(n) = C(S(n))$  (thus we always have  $C(n) < 10$ ).

**Solution.** For each case we must check the three defining properties of equivalence relations.

- i) a) Reflexivity: for any real function  $f$ ,  $f(0) = f(0)$ .
- b) Symmetry: if  $f(0) = g(0)$ , then also  $g(0) = f(0)$ .

**1.6.3. Relation on a set.** In the case when  $A = B$  we speak about a relation on the set  $A$ . We say that the relation  $R$  is:

- reflexive, if  $\text{id}_A \subseteq R$ , that is  $(a, a) \in R$  for every  $a \in A$ ,
- symmetric, if  $R^{-1} = R$ , that is if  $(a, b) \in R$ , then also  $(b, a) \in R$ ,
- antisymmetric, if  $R^{-1} \cap R \subseteq \text{id}_A$ , that is if  $(a, b) \in R$  and if also  $(b, a) \in R$ , then  $a = b$ ,
- transitive, if  $R \circ R \subseteq R$ , that is if  $(a, b) \in R$  and  $(b, c) \in R$  implies  $(a, c) \in R$ .

A relation is called an *equivalence relation* if it is reflexive, symmetric and transitive.



A relation is called an *ordering* if it is reflexive, transitive and antisymmetric. Orderings are usually denoted by the symbol  $\leq$ , that is the fact that element  $a$  is in relation with element  $b$  is written as  $a \leq b$ .

Notice that the relation  $<$ , that is “to be strictly smaller than”, is not an ordering on the set of real numbers, since it is not reflexive. The common practice (also used in this book) is to use the sign  $<$  or  $\subset$  for orderings, too.

A good example of an ordering is set inclusion. Consider the set  $2^A$  of all subsets of a finite set  $A$ . We have a relation  $\subseteq$  on the set  $2^A$  given by the property “being a subset”. Thus  $X \subseteq Z$  if  $X$  is a subset of  $Z$ . Clearly all three conditions from the definition of ordering are satisfied: if  $X \subseteq Y$  and  $Y \subseteq X$  then necessarily  $X$  and  $Y$  must be identical. If  $X \subseteq Y \subseteq Z$  then also  $X \subseteq Z$ , and reflexivity is clear from the definition. (As mentioned above, we shall mostly use the symbol  $\subset$  without excluding equality.)

We say that an ordering  $\leq$  on a set  $A$  is *complete*, if every two elements  $a, b \in A$  are *comparable*, that is, either  $a \leq b$  or  $b \leq a$ .

If  $A$  contains more than one element, there exist subsets  $X$  and  $Y$  where neither  $X \subseteq Y$  nor  $Y \subseteq X$ , so the ordering  $\subseteq$  is not complete on the set of all subsets of  $A$ .

The set of real numbers with the usual  $\leq$  is complete. Thus the subdomains  $\mathbb{N}$ ,  $\mathbb{Z}$ ,  $\mathbb{Q}$  come equipped with a complete ordering, too. On the other hand, there is no such natural complete ordering on  $\mathbb{C}$ . The absolute value provides a partial ordering there (comparing the radii of the circles in the complex plane).

**1.6.4. Partitions of an equivalence.** Every equivalence relation  $R$  on a set  $A$  defines also a *partition* of the set  $A$ , consisting of subsets of mutually equivalent elements, namely *equivalence classes*. For any  $a \in A$  we consider the set of elements, which are equivalent with  $a$ , that is

$$[a] = R_a = \{b \in A; (a, b) \in R\}.$$

Clearly  $a \in R_a$  by reflexivity. If  $(a, b) \in R$ , then  $R_a = R_b$  by symmetry and transitivity. Furthermore, if  $R_a \cap R_b \neq \emptyset$  then there is an element  $c$  in both  $R_a$  and  $R_b$  so that  $R_a = R_c = R_b$ . It follows that for every pair  $a, b$ , either  $R_a = R_b$ , or  $R_a$  and  $R_b$  are disjoint. That is, the equivalence classes are pairwise disjoint. Finally,  $A = \cup_{a \in A} R_a$ . That is, the set  $A$  is partitioned into equivalence classes. We write  $[a] = R_a$ ,



- c) Transitivity: if  $f(0) = g(0)$  and  $g(0) = h(0)$ , then also  $f(0) = h(0)$ . We conclude that the relation is an equivalence relation.
- ii) The relation is not reflexive, since for instance for the function  $f(x) = \sin(x)$  we have  $\sin(0) \neq \sin(1)$ . Moreover, it is not transitive.
- iii) The relation is neither reflexive (every line intersects itself), nor transitive.
- iv) This is an equivalence relation whose equivalence classes are well represented by the un-oriented lines through the origin.
- v) The relation clearly is not reflexive. Neither it is transitive.
- vi) In this case the answer is positive and we leave the verification of the details to the reader. □

**1.F.2.** Consider the set  $A = \{1, 2, 3, 4\}$  and let  $R$  be the relation over  $A$  defined by

$$R = \{(0, 0), (0, 1), (0, 3), (1, 0), (1, 1), (2, 3), (3, 3)\}$$

Show that  $R$  is *not* an equivalence relation by explaining which of the defining properties of an equivalence relation are not satisfied.

**Solution.** It is easy to see that  $R$  is not reflexive, since  $2 \in A$  but  $(2, 2) \notin R$ . It is not symmetric since for example  $(3, 2) \notin R$ . Finally it is not transitive since  $(1, 0)$  and  $(0, 3)$  they both belong to  $R$  but  $(1, 3) \notin R$ . □

**1.F.3.** Over the set  $A = \{a, b, c, d\}$  consider the relation  $R_1 = \{(a, a), (b, b), (c, c), (d, d), (b, a), (b, c), (b, d)\}$ . Is  $R_1$  an equivalence relation?

**Solution.** Since  $(k, k) \in R_1$  for any  $k \in A$ , the relation is reflexive. It is also transitive. However, it is not symmetric since for example  $(b, a) \in R_1$  but  $(a, b) \notin R_1$ . Hence  $R_1$  is not an equivalence relation. □

**1.F.4.** Verify the result of the previous task in Sage. ○

**1.F.5.** Let the relation  $R$  be defined over  $\mathbb{R}^2$  such that  $((a, b), (c, d)) \in R$  for arbitrary  $a, b, c, d \in \mathbb{R}$  if and only if  $b = d$ . Determine whether or not this is an equivalence relation. In the positive case, describe geometrically the partitioning determined by  $R$ . ○

**1.F.6.** Present the domain  $D$  and the codomain  $I$  of the relation  $R = \{(a, v), (b, x), (c, x), (c, u), (d, v), (f, y)\}$  defined between the sets  $A = \{a, b, c, d, e, f\}$  and  $B = \{x, y, u, v, w\}$ . Is the relation  $R$  a mapping? Next attempt to solve the problem using Sage. ○

**1.F.7.** Determine for each of the following relations over the set  $\{a, b, c, d\}$  whether it is an ordering and whether it is complete:

$$R_1 = \{(a, a), (b, b), (c, c), (d, d), (b, a), (b, c), (b, d)\},$$

$$R_2 = \{(a, a), (b, b), (c, c), (d, d), (d, a), (a, d)\},$$

$$R_3 = \{(a, a), (b, b), (c, c), (d, d), (a, b), (b, c), (b, d)\}.$$

and by the above, we can represent an equivalence class by any one of its elements.

**1.6.5. Existence of scalars.** As before, we assume to know what sets are, and indicate the construction of the natural numbers.



We denote the empty set by  $\emptyset$  (notice the difference between the symbol 0 for the zero and the empty set  $\emptyset$ ) and define

$$(1) \quad 0 := \emptyset, \quad n + 1 := n \cup \{n\},$$

in other words

$$0 := \emptyset, \quad 1 := \{0\}, \quad 2 := \{0, 1\}, \dots, \quad n + 1 := \{0, 1, \dots, n\}.$$

This notation says that if we have already defined the numbers  $0, 1, 2, \dots, n$ , then the number  $n + 1$  is defined as the set of all previous numbers.

We have defined the set of natural numbers  $\mathbb{N}$ .<sup>3</sup> Next, we should construct the operations  $+$  and  $\cdot$  and deduct their required properties. In order to do that in detail, we would have to pay more attention to basic understanding of sets. For example, once we know what a disjoint union of sets is, we may define the natural number  $c = a + b$  as the unique natural number  $c$  having the same number of elements as the disjoint union of  $a$  and  $b$ .

Of course, formally speaking, we need to explain what does it mean for two sets to have the same number of elements. Let us notice that in general, having the two sets  $A$  and  $B$  of the same “size” should mean that there exists a bijection  $A \rightarrow B$ . This is completely in accordance with our intuition for finite sets. However, it is much less intuitive with infinite sets. For example there is the same amount of all natural numbers and those with natural square roots (the bijection  $a \mapsto a^2$ ), although the picture in the solution to 1.A.3 could be read as “many of natural numbers do not have a rational square root”. We say, that each set which is bijective to natural numbers  $\mathbb{N}$  is *countable*. Sets bijective to some natural number  $n$  (as defined above) are called *finite* (with *number of elements*  $n$ ), while the sets which are neither finite nor countable are called *uncountable*.

So adding  $a + b$  is understood as increasing  $a$  iteratively by one exactly as many time as is the number of elements in  $b$ . Similarly, we multiply  $a \cdot b$  by adding  $a$  to zero as many times as is the number of elements in  $b$ . Now we should check the axioms for scalar operations for our operations  $+$  and  $\cdot$ . The (boring but tedious) check is left to the reader.

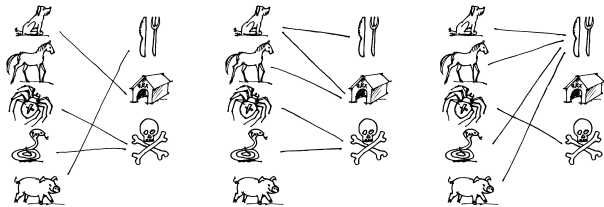
We can also define a relation  $\leq$  on  $\mathbb{N}$  as follows:  $m \leq n$ , if either  $m \in n$  or  $m = n$ . Clearly this is a complete ordering. For instance  $2 \leq 4$ , since

$$2 = \{\emptyset, \{\emptyset\}\} \in \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = 4.$$

<sup>3</sup>The concept of natural numbers based on the principle of “increasing by one” was known to all ancient civilisations, however they always had the smallest natural number one. The set theoretical approach was developed in 19th century and there zero got a logical smallest natural number as the counterpart of the empty set.

**Solution.** The relation  $R_1$  is an ordering, which is not complete (for instance neither  $(a, c) \notin R_1$  nor  $(c, a) \notin R_1$ ). The relation  $R_2$  is not anti-symmetric as it is both  $(a, d) \in R_2$  and  $(d, a) \in R_2$ , therefore it is not an ordering. The relations  $R_3$  does not define an ordering as well, since it is not transitive (for instance  $(a, b), (b, c) \in R_3$ , but  $(a, c) \notin R_3$ ).  $\square$

**1.F.8.** In the following three figures, icons are connected with lines such that people in different parts of the world could have assigned them. Determine whether the connection is a mapping, and whether it is injective, surjective or bijective.



**Solution.** In the first figure the connection is a mapping which is surjective but not injective, because both the snake and the spider are labeled as poisonous. The second figure is not a mapping but only a relation, since the dog is labeled both as a pet and as a meal. The third connection is again a mapping. This time it is neither injective nor surjective.  $\square$

**1.F.9.** Determine whether or not the mapping  $f$  is injective (one-to-one) or surjective (onto), when  $f$  is given by:

- (a)  $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}, \quad f((x, y)) = x + y - 10x^2;$
- (b)  $f : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}, \quad f(x) = (2x, x^2 + 10).$   $\circ$

**1.F.10.** Determine the number of equivalence relations over the set  $A = \{1, 2, 3, 4\}$ .

**Solution.** We divide the sought equivalences according to the types of corresponding partitions (given by number and cardinality of equivalence classes), and we count the number of partitions of a given type:

The type of partition	number of partitions of this type
1,1,1,1	1
2,1,1	$\binom{4}{2}$
2,2	$\frac{1}{2} \binom{4}{2}$
3,1	$\binom{4}{1}$
4	1

Thus in total we have 15 different equivalences.  $\square$

**1.F.11.** Determine the number of orderings of a four-element set.

**Solution.** We will consider all possible Hasse diagrams of orderings over a four-element set  $M$ . We count how many different orderings (recall that an ordering is a subset of a set  $M \times M$ ) the given Hasse diagram has. See the diagram:

In other words, the recurrent definition itself gives the relation  $n \leq n+1$ . and transitivity then gives  $n \leq k$  for all  $k$  obtained in this manner later.

This ordering of the positive integers or natural numbers (the number  $a$  is strictly smaller than  $b$  if  $a \in b$ ) has obviously got the following striking property: *every nonempty subset in  $\mathbb{N}$  or  $\mathbb{Z}^+$  has a smallest element.*

**1.6.6. Integers and rational numbers.** With the set  $\mathbb{N}$  of positive integers together with zero, we can always add two numbers together. Also, adding zero to a number does not change it. We can also define subtraction, but the result does not always belong to  $\mathbb{N}$ .

The basic idea of construction of the integers from the natural numbers or positive integers is to add to  $\mathbb{N}$  these missing results.

This can be done as follows: instead of subtraction, we will work with ordered pairs of numbers  $(a, b)$  representing the desired result  $a - b$ . It just remains to define which such pairs are equivalent (with respect to the result of subtraction).

The necessary relation is then:

$$(a, b) \sim (a', b') \iff a - b = a' - b' \iff a + b' = a' + b.$$

Note that the expression in the middle equation may not belong to  $\mathbb{N}$ , but the expression on the right always does. It is easy to check that it really is an equivalence, and we denote its classes as the integers  $\mathbb{Z}$ .

We define addition and subtraction on  $\mathbb{Z}$  using representatives. For instance

$$[(a, b)] + [(c, d)] = [(a + c, b + d)],$$

which is clearly independent of the choice of representatives.

It is always possible to choose a representative  $(a, 0)$  for natural numbers  $a$ , and a representative  $(0, a)$  for negative numbers  $-a$ . This is probably the simplest and easiest choice.

Next, we define multiplication of integers similarly to the addition (i.e., using representatives we aim at the representative of  $(a - b) \cdot (c - d)$ )

$$(a, b) \cdot (c, d) = (ac + bd, bc + ad).$$

This is clearly commutative and obviously does not depend on the choice of representatives.

Moreover, choosing positive or negative multiplicands with their special representatives as above, we get just the standard multiplication in  $\mathbb{N}$  taking the right care for the signs.

Now it is easy to see, that we have all the properties (CG1)–(CG4) and (R1)–(R4), see the paragraph 1.1.1. For multiplication, the neutral element is one, but for all numbers  $a$  other than zero and  $\pm 1$  there does not exist an integer  $a^{-1}$  with the property  $a \cdot a^{-1} = 1$ . Thus, for multiplication, we are missing the inverse elements. However, the property of the integral domain (ID) holds. This means that if the product of two integers equals zero, then at least one of them has to be zero.

We can construct the rational numbers  $\mathbb{Q}$  by adding all the missing multiplicative inverses by a method analogous to the construction of  $\mathbb{Z}$  from  $\mathbb{N}$ .

..... 1	!.. $\frac{4!}{2}$	!! $\frac{4!}{2}$	·^ $\frac{4!}{2}$	·v $\frac{4!}{2}$	·!  4!	N 4!	X $\frac{4!}{4}$
^ $\frac{4!}{6}$	v $\frac{4!}{6}$	!  4!	!  4!	Y $\frac{4!}{2}$	^ $\frac{4!}{2}$	◇ $\frac{4!}{2}$	!  4!

In total, there are 219 orderings over a four-element set. □

Many combinatorics problems involve relations, and examples can be found in the final section with additional material (see Section G).

On the set of all ordered pairs  $(p, q)$ ,  $q \neq 0$ , of integers, we define a relation  $\sim$  so that it models our expectation of the fractions  $p/q$ :

$$(p, q) \sim (p', q') \iff p/q = p'/q' \iff p \cdot q' = p' \cdot q.$$

Again, we are not able to formulate the expected behaviour in the middle equation when we work in  $\mathbb{Z}$ , but for the equation on the right this is indeed possible. This relation is a well-defined equivalence (think it through!).

If we formally write  $p/q$  instead of pairs  $(p, q)$ , we can define the operations of multiplication and addition by the well-known formulas

$$p/q \cdot r/s = pr/qs$$

$$p/q + r/s = ps/qs + qr/qs = (ps + qr)/qs.$$

**1.6.7. Remainder classes.** Another example of equivalence



classes is the remainder classes of integers. For a fixed natural number  $k$  we define an equivalence  $\sim_k$  so that two numbers  $a, b \in \mathbb{Z}$  are equivalent if they have the same remainder when divided by  $k$ . The resulting set of equivalence classes is denoted as  $\mathbb{Z}_k$ . This procedure is simplest for  $k = 2$ . This yields  $\mathbb{Z}_2 = \{[0], [1]\}$ , where zero stands for even numbers and one for odd numbers. It is easy to see that the addition and multiplication can be defined using representatives.

REMAINDER CLASSES RINGS AND FIELDS

**Theorem.** *The remainder class  $\mathbb{Z}_k$  is always a commutative ring of scalars. It is a commutative field of scalars (that is, the property (F) from the paragraph 1.1.1 is also satisfied) if and only if  $k$  is a prime.*

*If  $k$  is not prime, then  $\mathbb{Z}_k$  contains a divisor of zero, thus it is not an integral domain.*

**PROOF.** The second part is easy to see — if  $x \cdot y = k$  for natural numbers  $x, y$ , then the result of multiplying the corresponding classes  $[x] \cdot [y]$  is zero.

On the other hand, if  $x$  and  $k$  are relatively prime, then according to the Bezout equality, (which we derive later, see 11.1.2), there are natural numbers  $a$  and  $b$  satisfying

$$ax + bk = 1,$$

which for corresponding equivalence classes gives

$$[a] \cdot [x] + [0] = [a] \cdot [x] = [1]$$

and thus  $[a]$  is the inverse element to  $[x]$ . □

**G. Additional exercises for the whole chapter**

We will now present supplementary material related to the concepts discussed so far. We begin by reviewing some key equalities and inequalities that appear throughout the book, due to their versatility in various applications.

A) MATERIAL ON NUMBERS AND FUNCTIONS

**I.G.1. (Basic arithmetic and the triangular numbers)** Recall from high-school that the principle of mathematical induction can be used to prove the following in just a few steps:

$$\Delta_n := 1 + 2 + \dots + n = \frac{n(n+1)}{2}, \quad n \in \mathbb{Z}_+.$$

The numbers  $\Delta_n$  are called *triangle numbers*. Provide a shorter proof of this summation without using induction, and illustrate your method with a figure. Next, show that the number of distinct handshakes or wine glass clinks among  $n$  people equals the triangular number  $\Delta_{n-1}$ . Finally prove that  $\Delta_n - \Delta_{n-1} = n$ , and  $\Delta_n + \Delta_{n-1} = n^2$  for any  $n$ . ○

**I.G.2. (Average speed)** A bicyclist ascends a hill at 25km/h, and descends the same hill at 75km/h. Find the bicyclist's average speed for the entire trip, assuming the length of the hill is not needed for the calculation. ○

**1.G.3. The arithmetic–geometric mean inequality (AM-GM inequality).** (a) For any two non-negative real numbers  $a, b$  prove the inequality

$$\sqrt{ab} \leq \frac{a+b}{2}, \quad \text{or equivalently} \quad (ab)^{\frac{1}{2}} \leq \frac{a+b}{2}, \quad (\dagger)$$

with equality if and only if  $a = b$  (recall that  $\sqrt{ab}$  is called the *geometric mean* of  $a, b$ , while  $\frac{a+b}{2}$  is the *arithmetic mean* of  $a, b$ ). Next use  $(\dagger)$  to prove that any four non-negative real numbers  $a, b, c, d$ , the following inequality holds:

$$(abcd)^{\frac{1}{4}} \leq \frac{a+b+c+d}{4}, \quad (\ddagger)$$

with equality if and only if  $a = b = c = d$ .

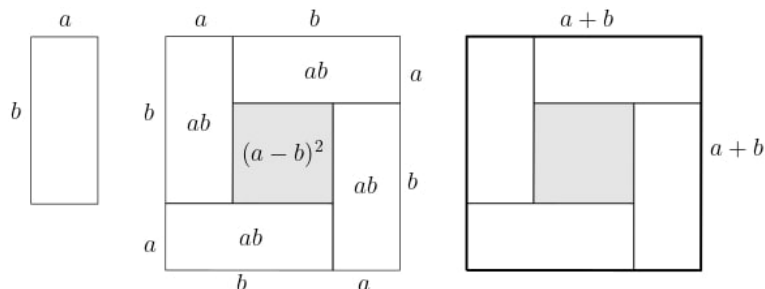
(b) Present a proof using mathematical induction for the general arithmetic–geometric mean inequality

$$\sqrt[n]{a_1 \cdot a_2 \cdot \dots \cdot a_n} \leq \frac{a_1 + a_2 + \dots + a_n}{n},$$

with equality if and only if  $a_1 = a_2 = \dots = a_n$ , where  $a_i$  are non-negative real numbers for all  $i = 1, \dots, n$ .

**Solution.** (a) There are numerous alternative proofs for the AM-GM inequality. For instance, take the square in both sides of the relation given in  $(\dagger)$ . We obtain  $ab \leq (a+b)^2/4$ , which simplifies to  $4ab \leq a^2 + 2ab + b^2$ . This is equivalent to stating that  $(a-b)^2 \geq 0$ . This last inequality is obviously true, as the square of a real number is non-negative. Moreover, equality holds if and only if  $a = b$ . Since all steps are reversible, the statement is proven.

A more geometric proof can be described using the figure below. This illustrates four rectangles, each of area  $ab$ , placed inside a square of area  $(a+b)^2$ . We observe that  $4ab$  does not exceed  $(a+b)^2$ , with the “wasted” area being the shaded region in the figure. By demonstrating that the shaded area is  $(a-b)^2$  we can conclude that equality holds if and only if  $a = b$ .



The second inequality follows directly from  $(\dagger)$ , when combined with the known properties of powers of real numbers:

$$(abcd)^{\frac{1}{4}} = \left( (abcd)^{\frac{1}{2}} \right)^{\frac{1}{2}} = \left( \sqrt{ab} \sqrt{cd} \right)^{\frac{1}{2}} \stackrel{(\dagger)}{\leq} \frac{\sqrt{ab} + \sqrt{cd}}{2} \stackrel{(\dagger)}{\leq} \frac{\frac{a+b}{2} + \frac{c+d}{2}}{2} = \frac{a+b+c+d}{4}.$$

(b) To simplify the notation, suppose that  $\mathcal{Y}(n)$  stands for the hypothesis that the given inequality is valid. Above we proved the result for  $n = 2, 4$ , that is, we have confirmed  $\mathcal{Y}(2)$  and  $\mathcal{Y}(4)$ , and in a similar way one can show that the result is valid for  $n = 8$ , that is,

$$(a_1 \cdot a_2 \cdot \dots \cdot a_8)^{\frac{1}{8}} = \left( (a_1 \cdot \dots \cdot a_4)^{\frac{1}{4}} (a_5 \cdot \dots \cdot a_8)^{\frac{1}{4}} \right)^{\frac{1}{2}} \stackrel{(\dagger)}{\leq} \frac{(a_1 \cdot \dots \cdot a_4)^{\frac{1}{4}} + (a_5 \cdot \dots \cdot a_8)^{\frac{1}{4}}}{2} \stackrel{(\dagger)}{\leq} \frac{a_1 + \dots + a_8}{8}.$$

Hence we have confirmed  $\Upsilon(2)$ ,  $\Upsilon(4)$  and  $\Upsilon(8)$ , and one can obtain the statement for any  $2^k$  with  $k \geq 1$ , that is, verify  $\Upsilon(2^k)$  for all integers  $k \geq 1$  (we leave this verification to the reader). Let us now explain how  $\Upsilon(4)$  can be used to derive  $\Upsilon(3)$ . Set

$$\mu := \frac{a_1 + a_2 + a_3}{3}$$

for the arithmetic mean of some real positive numbers  $a_1, a_2, a_3$ . A crucial but simple observation is that  $\mu$  is also the arithmetic mean of the numbers  $a_1, a_2, a_3, \mu$ , that is,

$$\frac{a_1 + a_2 + a_3}{3} = \frac{a_1 + a_2 + a_3 + \mu}{4},$$

or in other words  $\mu = \frac{a_1 + a_2 + a_3 + \mu}{4}$ . A proof of this is easy: By definition we have  $3\mu = a_1 + a_2 + a_3$ , or equivalently  $4\mu = a_1 + a_2 + a_3 + \mu$ . Then a division by 4 gives the claim. Now we can apply  $\Upsilon(4)$  on the numbers  $a_1, a_2, a_3, \mu$ :

$$\sqrt[4]{(a_1 a_2 a_3 \mu)} = (a_1 a_2 a_3 \mu)^{\frac{1}{4}} \leq \frac{a_1 + a_2 + a_3 + \mu}{4} = \mu.$$

Considering the fourth power of both sides in this inequality, we obtain  $abc\mu \leq \mu^4$  which can be rewritten as  $a_1 a_2 a_3 \leq \mu^3$ . Thus, after taking cube roots in both sides we arrive to the result, i.e.,  $(a_1 a_2 a_3)^{\frac{1}{3}} \leq \mu$ . Before describe the general step, keep in mind that  $a_1 a_2 a_3 \mu$  can be rewritten as  $a_1 a_2 a_3 \mu^{2^2-3}$  and  $a_1 + a_2 + a_3 + \mu = a_1 + a_2 + a_3 + (2^2 - 3)\mu$ , since  $2^2 - 3 = 1$ .

Now one can use  $\Upsilon(8)$ , to obtain also  $\Upsilon(5)$ ,  $\Upsilon(6)$ , and  $\Upsilon(7)$ , and so forth. However, let us explicitly present the general argument. Observe that in order to compute  $\Upsilon(3)$  from  $\Upsilon(4)$ , we extended the triple  $(a_1, a_2, a_3)$  into the quadruple  $(a_1, a_2, a_3, \mu) = (\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4)$ , so that we could apply  $\Upsilon(4)$ . We will use the analogous technique to prove  $\Upsilon(n)$  for all  $n < 2^k$ . Hence, the goal is to extend  $(a_1, a_2, \dots, a_n)$  to  $(\hat{a}_1, \dots, \hat{a}_{2^k})$ , so that we can apply  $\Upsilon(2^k)$  on the latter set. A question that may arise is how to extend this result, and the case  $n = 3$  provides the necessary motivation. Indeed, consider the following:

$$\hat{a}_j := \begin{cases} a_j, & j = 1, 2, \dots, n, \\ \mu := \frac{a_1 + a_2 + \dots + a_n}{n}, & j = n + 1, \dots, 2^k \end{cases}$$

(for  $n = 3$  this formula yields the expressions used above, and in particular we have  $k = 2$  for this case). Apply now  $\Upsilon(2^k)$  to  $(\hat{a}_1, \dots, \hat{a}_{2^k})$ . This gives

$$(a_1 \cdots a_n \mu^{2^k-n})^{\frac{1}{2^k}} \leq \frac{a_1 + \cdots + a_n + (2^k - n)\mu}{2^k} = \mu.$$

Obviously,  $(\mu^{2^k-n})^{\frac{1}{2^k}} = \mu \mu^{-\frac{n}{2^k}}$ , therefore the previous formula is also written as

$$(a_1 \cdots a_n \mu^{2^k-n})^{\frac{1}{2^k}} = (a_1 \cdots a_n)^{\frac{1}{2^k}} \mu \mu^{-\frac{n}{2^k}} \leq \mu.$$

Thus we can eliminate  $\mu$  from both sides, i.e., we have proved till now that  $(a_1 \cdots a_n)^{\frac{1}{2^k}} \mu^{-\frac{n}{2^k}} \leq 1$ . Raising both sides of this inequality to the power of  $\frac{2^k}{n}$ , yields the result:

$$(a_1 \cdots a_n)^{\frac{1}{n}} \mu^{-1} \leq 1, \quad \text{or equivalent} \quad (a_1 \cdots a_n)^{\frac{1}{n}} \leq \mu.$$

This completes the main proof, and the equality case can be easily derived. □

*The proof given above goes back almost two centuries, and is due to the French mathematician A.-L. Cauchy (1789–1856).*



*Note that this is a special kind of mathematical induction, often referred to as the “forward-backward” induction.*

*This is because we first derived the result for all the powers of 2, and next proceeded with another induction to establish it for all positive integers. As we pronounced, the AM-GM inequality admits many different proofs, most of them based on induction. There are also methods arising in term of calculus, which allow the presentation of shorter proofs. For example, much later in Chapter 8, we will provide a proof of the AM-GM inequality in terms of the so-called Lagrange multipliers.*

**1.G.4. Bernoulli inequality.** For any  $x \in \mathbb{R}$  with  $x > -1$  based on mathematical induction one can easily prove that  $(1+x)^n \geq 1+nx$  for any  $n \in \mathbb{N}$ , see also 5.4.1. This is the so called *Bernoulli inequality*. Prove that the Bernoulli inequality yields the AM-GM inequality.

**Solution.** For  $n = 1$  we have nothing to prove, so we may assume that  $n \geq 2$ . Setting  $y = x + 1$  with  $x > -1$ , so that  $y > 0$ , the Bernoulli inequality takes the form

$$y^n \geq 1 + n(y - 1). \quad (*)$$

Now, to simplify the description let us set  $A(n) = \frac{a_1 + \dots + a_n}{n}$  and  $G(n) = (a_1 \dots a_n)^{\frac{1}{n}}$ , so that the AM-GM inequality simply reads by  $G(n) \leq A(n)$ . Obviously  $\frac{A(n)}{A(n-1)} > 0$  and as  $y$  we may consider the ratio  $\frac{A(n)}{A(n-1)}$  with  $n \geq 2$ . Then, using the relation (\*) we deduce that

$$\left(\frac{A(n)}{A(n-1)}\right)^n \geq 1 + n \left(\frac{A(n)}{A(n-1)} - 1\right) = \frac{A(n-1) + nA(n) - nA(n-1)}{A(n-1)} = \frac{a_n}{A(n-1)},$$

and this can be equivalently rephrased as  $A^n(n) \geq a_n A^{n-1}(n-1)$ . Using this inequality successively we obtain

$$\begin{aligned} A^n(n) &\geq a_n A^{n-1}(n-1) \geq a_n \cdot a_{n-1} A^{n-2}(n-2) \geq \dots \geq a_n \cdot a_{n-1} \cdot \dots \cdot a_2 A^1(1) \\ &= a_n \cdot a_{n-1} \cdot \dots \cdot a_2 \cdot a_1 = G^n(n), \end{aligned}$$

that is,  $A^n(n) \geq G^n(n)$ , for any  $n \geq 2$ . Raising both sides to the power of  $\frac{1}{n}$  we obtain the desired  $A(n) \geq G(n)$ .  $\square$

**1.G.5. Symbolic computations with Sage and polynomials.** In SageMath, for a series of applications as doing symbolic computations, solving equations, plotting functions, etc. are available the so called *symbolic variables*. Symbolic variables are often used to represent "indeterminates," which are variables that do not have a fixed value. This concept is frequently employed in mathematical problem-solving to allow for general solutions and manipulations, such as simplifying expressions, solving equations, and performing symbolic differentiation and integration (see Chapter 5 and Chapter 6, respectively). By using indeterminates, we can explore mathematical relationships and derive formulas that hold true for any value of the variable, providing a powerful tool for both theoretical and applied mathematics.

On the other hand, polynomials form a fundamental component of algebra and calculus, and symbolic computation provides a powerful toolset for examining their properties and behavior. Let us explain how using symbolic variables, we can manipulate polynomials easily. First, to establish a symbolic variable we type `t = var("t")`, or simply `var("t")`, or

```
t=SR.var("x"); t
```

which returns a symbolic variable, named `t` by Sage. We stress that Sage does this *automatically* for the variable `x`, but not for other variables. Thus, given the cell

```
y=var("y"); x-y
```

Sage will understand `x - y` as an expression of two variables `x, y`. In the opposite, typing directly `x - y` an error will be printed out, declaring that the name "y" is not defined. It is also possible to create more than one symbolic variable, in a row. This for example allows to use expressions depending either on many variables, or both on variables and parameters, as in the examples below:

```
x, y, z=var("x, y, z"); x+y+z
t, a, b, c=var("t, a, b, c"); a*t**2+b*t+c
```

The cell

```
var("a, b, c"); f=a*x**2+b*x+c; print([f.coefficient(x**2), f.coefficient(x)])
```

returns the coefficients of the terms  $x^2$  and  $x$ , respectively, as `[a, b]`, which can be useful in complicated expressions. Of course, if you want to solve the equation  $ax^2 + bx + c = 0$  with  $a > 0$ , then type

```
x, a, b, c=var("x, a, b, c"); assume(a>0); show(solve(a*x**2+b*x+c==0, x))
```

In this case Sage returns the known solutions in terms of `a, b, c`, check the output yourselves. An alternative to introduce symbolic variables is the command `x = SR.var("x", n)`. This creates the symbolic variables  $x_0, \dots, x_{n-1}$  for some positive integer `n`, and each of them can be accessed by typing `x[i]`, for some  $i \in \{0, \dots, n-1\}$ . As an example, type

```
x=SR.var("x"); a=SR.var("a", 4); a[0]+a[1]*x+a[2]*x**2+a[3]*x**3
```

Executing this cell we obtain  $a_3 * x^3 + a_2 * x^2 + a_1 * x + a_0$ , hence fixing the parameters  $a_3, a_2, a_1, a_0$  we can define a real polynomial of degree three. For such substitutions one may apply the following code:

```
x=SR.var("x"); a=SR.var("a", 4)
a[0]+a[1]*x+a[2]*x**2+a[3]*x**3
f=a[0]+a[1]*x+a[2]*x**2+a[3]*x**3
f.subs(a[0]==6, a[1]==4, a[2]==2, a[3]==1/8)
```

This prints out the polynomial  $\frac{1}{8} * x^3 + 2 * x^2 + 4 * x + 6$ , which we may plot by typing (produce the figure in your editor)

```
plot(1/8*x^3 + 2*x^2 + 4*x + 6, (x, -50, 50))
```

Notice to factor a polynomial we may use the command `factor`:

```
p(x)=x^5 + x^4 - 8*x^3 + 11*x^2 - 15*x + 2; show(factor(p(x)))
```

In this case Sage prints out the expression  $(x^4 + 3x^3 - 2x^2 + 7x - 1)(x - 2)$ . We will meet further option manipulating polynomials in Sage in the sequel. We will encounter more options for manipulating polynomials in SageMath later in this text.

**1.G.6.** Suppose that  $P(x)$  is a polynomial (with constant coefficients) satisfying the following: divided by  $x + 2$  leaves remainder  $-1$ , divided by  $x + 3$  leaves remainder  $1$ , while divided by  $x - 3$  leaves remainder  $4$ . Find the remainder  $v(x)$  of the division of  $P(x)$  by the polynomial  $(x + 2)(x + 3)(x - 3)$ .

**Solution.** Let us write  $P(x) = q(x) \cdot \pi(x) + v(x)$ , where  $q(x) = (x + 2)(x + 3)(x - 3)$  is the divisor,  $\pi(x)$  is the quotient and  $v(x)$  the remainder. Since the divisor is of degree  $3$ , the remainder  $v(x)$  is a polynomial at most of degree  $2$ , say  $v(x) = ax^2 + bx + c$ . By assumption we have  $P(-2) = -1$ ,  $P(-3) = 1$  and  $P(3) = 4$ . Since  $P(x) = (x_2)(x + 3)(x - 3) \cdot \pi(x) + ax^2 + bx + c$ , we get the following system of equations:

$$\{4a - 2b + c = -1, \quad 9a - 3b + c = 1, \quad 9a + 3b + c = 4\}.$$

Solving this system we obtain  $a = 1/2 = b$  and  $c = -2$ . For verifying this, use for example Sage via the block

```
var("a, b, c"); f1=4*a-2*b+c+1; f2=9*a-3*b+c-1; f3=9*a+3*b+c-4
solve((f1==0, f2==0, f3==0), a,b,c)
```

□

**1.G.7.** Find a polynomial  $P(x)$  of fourth degree satisfying  $P(1) = 1$ , such that  $P(x+1)$  is divided by  $(x-1)^2$ , and  $P(x-1)$  is divided by  $(x+1)^2$ .

**Solution.** By assumption, we have

$$P(x+1) = (x-1)^2 \cdot \pi_1(x), \quad \text{and} \quad P(x-1) = (x+1)^2 \cdot \pi_2(x),$$

where  $\pi_1(x), \pi_2(x)$  are the corresponding quotients. Using these equations, one can obtain the relations

$$P(x) = (x-2)^2 \cdot \pi_1(x-1), \quad \text{and} \quad P(x) = (x+2)^2 \cdot \pi_2(x+1),$$

respectively. This means that  $P(x)$  is divided by  $(x-2)^2$  and also by  $(x+2)^2$ . But then  $P(x)$  will be divided also by the product  $(x-2)^2(x+2)^2 = (x^2-4)^2$ .<sup>13</sup> Now,  $P(x)$  is of degree  $4$ , hence the previous conclusion implies that  $P(x) = c \cdot (x^2-4)^2$ , for some constant  $c \in \mathbb{R} \setminus \{0\}$ . Based on the condition  $P(1) = 1$  we compute  $c = 1/9$ , and hence  $P(x) = \frac{1}{9}(x^2-4)^2$ . □

**1.G.8.** Given the polynomial  $P(x) = x^2 + x + 2$ , show that the polynomial  $Q(x) := P(P(x)) - P(x) - 3$  is divided by  $P(x) - 1$ . Which is the quotient? ○

**1.G.9.** Prove that the polynomial  $P(x)$  of degree  $4$  satisfying  $P(0) = 0$  and  $P(x+1) - P(x) = 4x^3$ , admits  $x_0 = 1$  as a double root. ○

**1.G.10.** (a) Factor the polynomial  $p(x) = 4x^4 - 8x^3 - 3x^2 + 7x - 2$  using Horner's scheme and indicate the multiplicities of its roots. Verify your answer via Sage and present the graph of  $p(x)$  for  $x \in [-2, 2]$ .

(b) Consider the polynomials  $p_1(x) = x^4 + 2x^3 + x^2 + x + 12$ ,  $p_2(x) = 2x^5 + x^4 - 53x^3 - 7x + 3$ ,  $q_1(x) = x + 2$  and  $q_2(x) = x^2 - 5x + 1$ . Show that the Euclidean divisions of  $p_i(x)$  by  $q_i(x)$  for  $i = 1, 2$ , satisfy  $v_2(x) = -62x + v_1(x)$  and moreover  $\pi_2(x) = 2\pi_1(x) + 11x^2 - 2x - 9$ , where we write  $p_i(x) = q_i(x)\pi_i(x) + v_i(x)$  with  $i = 1, 2$ .

**Solution.** (a) Dividing the factors of  $2$  by all factors of  $4$  ones gets a list of all possible rational roots. In particular, we see that  $\rho_1 = -1$  is a root,  $p(\rho_1) = 4 + 8 - 3 - 7 - 2 = 0$ . Thus, using the Horner's table we get

$$\begin{array}{r|rrrrr} \rho_1 = -1 & 4 & -8 & -3 & 7 & -2 \\ & & -4 & 12 & -9 & 2 \\ \hline & 4 & -12 & 9 & -2 & 0 \end{array}$$

Hence  $p(x) = (x - \rho_1)(4x^3 - 12x^2 + 9x - 2) = (x + 1)(4x^3 - 12x^2 + 9x - 2)$ . Next we see that  $\rho_2 = 2$  is a root of  $4x^3 - 12x^2 + 9x - 2$ , thus using the Horner's method for this polynomial we get

$$\begin{array}{r|rrrr} \rho_2 = 2 & 4 & -12 & 9 & -2 \\ & & 8 & -8 & 2 \\ \hline & 4 & -4 & 1 & 0 \end{array}$$

<sup>13</sup>Here we are based on the general fact that if a polynomial  $P(x)$  (with constant coefficients) is divided by the monomials  $(x - \rho_1), (x - \rho_2), \dots, (x - \rho_k)$ , where  $\rho_i \neq \rho_j$  for  $1 \leq i \neq j \leq k$ , then  $P(x)$  is also divided by the product  $(x - \rho_1) \cdot (x - \rho_2) \cdot \dots \cdot (x - \rho_k)$ . The converse is also true.

Thus  $p(x) = (x - \rho_1)(x - \rho_2)(4x^2 - 4x + 1) = (x + 1)(x - 2)(2x - 1)^2$ , or equivalently  $p(x) = 4(x - \rho_1)(x - \rho_2)(x - \rho_3)^2$  where  $\rho_1 = -1$ ,  $\rho_2 = 2$  and  $\rho_3 = \frac{1}{2}$ . Note that  $\rho_3$  is a double root. In Sage we simply type

```
p(x)=4*x^4-8*x^3-3*x^2+7*x-2; factor(p(x))
```

which returns  $(2 * x - 1)^2 * (x + 1) * (x - 2)$ . In this block one can specify the ring over which the factorization is implemented by typing  $x = \mathbb{QQ}[x]$  at the beginning of the code. This defines  $x$  as a polynomial over the rational numbers. However, in our case, this approach will yield the same factorization result as the default setting. Additionally, by typing  $p.roots()$  SageMath will return the roots of the polynomial  $p$ , along with their multiplicities:

```
[(2, 1), (-1, 1), (1/2, 2)] # in the pair (a, b), a denotes a root, b denotes its multiplicity
```

As for the graph of  $p(x)$  for  $x \in [-2, 2]$ , one may add the syntax

```
plot(4*x^4-8*x^3-3*x^2+7*x-2, x, -2, 2)
```

(b) The first division gives

$$\begin{array}{r} x^3 + x - 1 \\ x + 2 \overline{) x^4 + 2x^3 + x^2 + x + 12} \\ \underline{-x^4 - 2x^3} \phantom{+ x^2 +} \\ x^2 + x \phantom{+ 12} \\ \underline{-x^2 - 2x} \phantom{+ 12} \\ -x + 12 \\ \phantom{-x +} \underline{x + 2} \\ 14 \end{array}$$

Thus  $p_1(x) = q_1(x)\pi_1(x) + v_1(x)$ , where  $\pi_1(x) = x^3 + x + 1$  and  $v_1(x) = 14$  (constant). For the second division one gets

$$\begin{array}{r} 2x^3 + 11x^2 - 11 \\ x^2 - 5x + 1 \overline{) 2x^5 + x^4 - 53x^3 - 7x + 3} \\ \underline{-2x^5 + 10x^4 - 2x^3} \phantom{- 7x + 3} \\ 11x^4 - 55x^3 \phantom{- 7x + 3} \\ \underline{-11x^4 + 55x^3 - 11x^2} \phantom{- 7x + 3} \\ -11x^2 - 7x + 3 \\ \phantom{-11x^2 -} \underline{11x^2 - 55x + 11} \\ -62x + 14 \end{array}$$

hence  $q_2(x) = q_2(x)\pi_2(x) + v_2(x)$  with  $\pi_2(x) = 2x^3 + 11x^2 - 11$  and  $v_2(x) = -62x + 14$ , respectively. Thus  $v_2(x) = -62x + v_1(x)$  and the second relation follows since

$$\begin{array}{r} 2 \\ x^3 + x - 1 \overline{) 2x^3 + 11x^2 - 11} \\ \underline{-2x^3} \phantom{- 11} \\ 11x^2 - 2x - 9 \end{array}$$

□

**1.G.11. (Polynomial division via SageMath)** Given a polynomial  $p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$  and a number  $x_0$ , use Sage to implement the division of  $p(x)$  by  $(x - x_0)$ , resulting in a quotient polynomial  $q(x)$  and a remainder  $r$ . Next use your program to confirm that dividing  $p(x) = 4 - 3x + 5x^2 - 2x^3 + x^4$  by  $(x - 3)$  results in the quotient  $q(x) = 21 + 8x + x^2 + x^3$  and the remainder  $r = 67$ . ○

*The omission of basic features of integers, such as their divisibility and the theory of prime numbers, up to this point, has been intentional. It is more practical to explore these topics in detail in Chapter 11, which is entirely dedicated to number theory. For this reason, we will delay discussing additional tasks related to the beautiful theory of integers and prime numbers, and instead, focus on problems involving complex numbers and their geometry for now.*

**1.G.12. (True or False?)** Answer the following True or False questions, providing a proof to support your statements.

(1) “The expression of the complex number  $z = \frac{1 + 2i}{4 + 5i}$  in the form  $z = x + iy$  is given by  $-\frac{6}{41} + i\frac{13}{41}$ ”.



- (2) “The ratio  $\frac{(1+i)^{2023}}{(1-i)^{2020}}$  equals  $2i - 2$ ”.
- (3) “The product  $zw$  of two elements  $z, w \in S^1$  on the unit circle  $S^1 = \{z \in \mathbb{C} : |z| = 1\}$  belongs also to  $S^1$ ”.
- (4) “Any element  $z \in S^1$  has an inverse which also belongs to  $S^1$ ”.
- (5) “The cubic equation  $z^3 + i = 0$  has three solutions over  $\mathbb{C}$ , given by  $z_1 = i, z_2 = \frac{\sqrt{3}}{2} - i\frac{1}{2}, z_3 = -\frac{\sqrt{3}}{2} - i\frac{1}{2}$ .”

○

**I.G.13.** Consider the complex numbers  $z_1 = 2 + 6i$  and  $z_2 = -4 - 2i$ . Suppose that there is another complex number  $w = x + iy$  which is at a distance of  $2\sqrt{2}$  from  $z_2$ . Determine the positions of  $w$  on the plane  $\mathbb{R}^2$ , such that the triangle formed by the the points  $P_1 = [2, 6], P_2 = [-4, -2]$  and  $P_3 = [x, y]$  (representing  $z_1, z_2$  and  $w$ , respectively) is a right triangle. In particular, show that there are two possible solutions for  $w$ , and illustrate both of them using SageMath. ○

**1.G.14.** Let  $a, b, c, d \in \mathbb{R}$  with  $ac > 0$  and  $ad = bc$ . Show that the polynomial  $P(x) = ax^3 + bx^2 + cx + d$  has a unique real solution and two purely imaginary solutions.

**Solution.** The condition  $ac > 0$  implies that  $a \neq 0$ , and so  $d = \frac{bc}{a}$ . Therefore we see that

$$P(x) = ax^3 + bx^2 + cx + \frac{bc}{a} = \frac{1}{a}(a^2x^3 + abx^2 + acx + bc) = \frac{1}{a}(ax^2(ax + b) + c(ax + b)) = \frac{1}{a}(ax + b)(ax^2 + c).$$

As a consequence, the equation  $P(x) = 0$  is equivalent to  $(ax + b)(ax^2 + c) = 0$ , that is,  $x = -\frac{b}{a}$  or  $x^2 = -\frac{c}{a} < 0$ . So we have a real solution  $x_0 = -\frac{b}{a}$  and two purely imaginary given by  $x_{1,2} = \pm i\sqrt{\frac{c}{a}}$ . □

*Euler's identity for complex numbers states that any complex number  $z = r(\cos(\varphi) + i \sin(\varphi))$ , where  $r = |z|$  and  $\varphi$  is the argument of  $z$ , can be expressed as  $z = r e^{i\varphi}$ . Euler's identity thus provides an alternative to the polar form of  $z$ . For example, given  $z \in S^1$  we get  $z = \cos(\varphi) + i \sin(\varphi) = e^{i\varphi}$ , such that  $|z| = 1$ . It is worth noting that the exponential  $e^{iy}$  of an imaginary number  $iy$  is given by the power series  $\sum_{n=0}^{\infty} \frac{(iy)^n}{n!}$ . We will present a proof of this claim later, in Section 5.4.12, which discusses the relationship between exponentials and trigonometric functions. However, it is useful to get an initial understanding of this concept at this point.*



**1.G.15.** Based on the Euler's identity, provide an answer to the following tasks:

- (a) Express  $e^{i\pi/2}, e^{i\pi}$ , and  $e^{i2\pi}$  in algebraic form.
- (b) Show that  $\cos(\varphi) = \frac{e^{i\varphi} + e^{-i\varphi}}{2}$  and  $\sin(\varphi) = \frac{e^{i\varphi} - e^{-i\varphi}}{2i}$ , respectively. Deduce that  $\tan(\varphi) = -i \frac{e^{i\varphi} - e^{-i\varphi}}{e^{i\varphi} + e^{-i\varphi}}$ .
- (c) Show that  $|e^{i\varphi}| = 1$ .
- (d) Using the formula  $e^{i(\alpha+\beta)} = e^{i\alpha}e^{i\beta}$ , prove the following classical trigonometric identities:

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta), \quad \sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta).$$

**Solution.** (a) Using Euler's identity we obtain

$$e^{i\pi/2} = \cos(\pi/2) + i \sin(\pi/2) = i = \sqrt{-1}, \quad e^{i\pi} = \cos(\pi) + i \sin(\pi) = -1, \quad e^{i2\pi} = \cos(2\pi) + i \sin(2\pi) = 1.$$

(b) We have  $e^{i\varphi} = \cos(\varphi) + i \sin(\varphi)$  and  $\cos(\varphi) - i \sin(\varphi) = e^{-i\varphi} = \overline{e^{i\varphi}}$  (see also 5.4.12). By adding these two equations and solving for  $\cos(\varphi)$ , we derive the formula for  $\cos(\varphi)$ . In a similar way we obtain the formula for  $\sin(\varphi)$ , and the final claim follows in combination with the usual definition of the trigonometric function  $\tan$ , that is,  $\tan(\varphi) := \frac{\sin(\varphi)}{\cos(\varphi)}$ .

(c) Indeed,  $|e^{i\varphi}| = |\cos(\varphi) + i \sin(\varphi)| = \sqrt{\cos^2(\varphi) + \sin^2(\varphi)} = \sqrt{1} = 1$ .

(d) We see that

$$e^{i(\alpha+\beta)} = e^{i\alpha}e^{i\beta} = (\cos(\alpha) + i \sin(\alpha)) \cdot (\cos(\beta) + i \sin(\beta)) = (\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)) + i(\sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)).$$

On the other hand  $e^{i(\alpha+\beta)} = \cos(\alpha + \beta) + i \sin(\alpha + \beta)$ . By comparing the real/imaginary parts of these two equations we obtain the given formulas. □

**1.G.16.** (a) Show that the following equation does not admit a real solution:

$$(1 + iz)^{2025} = \frac{2025 + i}{1 + i2025}, \quad z \in \mathbb{C}. \quad (*)$$

(b) Consider the complex number  $w = \frac{1}{2} \left( z + \frac{1}{z} \right)$  with  $z \in \mathbb{C}$  such that  $\arg(z) = \varphi \neq k\pi$  for all  $k \in \mathbb{Z}$  and  $\text{Im}(w) = 0$ . Show that the  $|z| = 1$ , i.e.,  $z$  lies on the unit circle.

**Solution.** (a) Suppose in the opposite that the equation (\*) has a real solution  $z = x \in \mathbb{R}$ , that is,  $(1 + ix)^{2025} = \frac{2025 + i}{1 + 2025i}$ .

Taking absolute values in both sides we have the following equivalent relations:

$$\begin{aligned} |(1 + ix)^{2025}| = \left| \frac{2025 + i}{1 + i2025} \right| &\iff |(1 + ix)|^{2025} = \frac{|2025 + i|}{|1 + i2025|} \iff (\sqrt{1 + x^2})^{2025} = \frac{\sqrt{2025^2 + 1}}{\sqrt{1 + 2025^2}} \iff \\ (\sqrt{1 + x^2})^{2025} = 1 &\iff \sqrt{1 + x^2} = 1 \iff 1 + x^2 = 1 \iff x = 0. \end{aligned}$$

However,  $x = 0$  cannot be a solution of the equation (\*) as it follows from the computation below:

$$1^{2025} = \frac{2025 + i}{1 + i2025} \iff 1 + i2025 = 2025 + i,$$

a contradiction. Hence the given equation does not admit a real solution.

(b) By assumption,  $\varphi$  is the (principle) argument of  $z$  and we may use the polar form  $z = |z|(\cos \varphi + i \sin \varphi)$ . Thus we can write

$$\begin{aligned} w &= \frac{1}{2} \left( |z|(\cos \varphi + i \sin \varphi) + \frac{1}{|z|(\cos \varphi + i \sin \varphi)} \right) = \frac{1}{2} \left( |z|(\cos \varphi + i \sin \varphi) + \frac{1}{|z|}(\cos(-\varphi) + i \sin(-\varphi)) \right) \\ &= \frac{1}{2} \left( |z|(\cos \varphi + i \sin \varphi) + \frac{1}{|z|}(\cos(\varphi) - i \sin(\varphi)) \right) = \frac{1}{2} \left( \left|z| + \frac{1}{|z|} \right) \cos \varphi + i \frac{1}{2} \left( |z| - \frac{1}{|z|} \right) \sin \varphi. \end{aligned}$$

Here we have used that the complex number  $e^{i\varphi} = \cos \varphi + i \sin \varphi$  satisfies (see also the previous problem)

$$\frac{1}{e^{i\varphi}} = e^{-i\varphi} = \cos(-\varphi) + i \sin(-\varphi) = \cos(\varphi) - i \sin(\varphi).$$

Thus now the requirement  $\text{Im}(w) = 0$  (in combination with the condition  $\varphi \neq k\pi$ ) gives

$$\frac{1}{2} \left( |z| - \frac{1}{|z|} \right) \sin \varphi = 0 \iff |z| - \frac{1}{|z|} = 0 \iff |z|^2 = 1 \iff |z| = 1.$$

□

**1.G.17. Roots of unity and regular  $n$ -polygons.** (a) Prove that the cube roots of 1 are the vertices of an equilateral triangle inscribed in the unit circle  $S^1$  (on the plane), with side length given by  $\sqrt{3}$ . Verify your computation via Sage.

(b) As a generalization, deduce that for any positive integer  $n$  the equation  $z^n = 1$ , where  $z \in \mathbb{C}$ , has  $n$ -roots  $z_1, z_2, \dots, z_n$  called the  $n$ th roots of unity. Prove that these solutions form the vertices of the regular  $n$ -polygon in the plane.

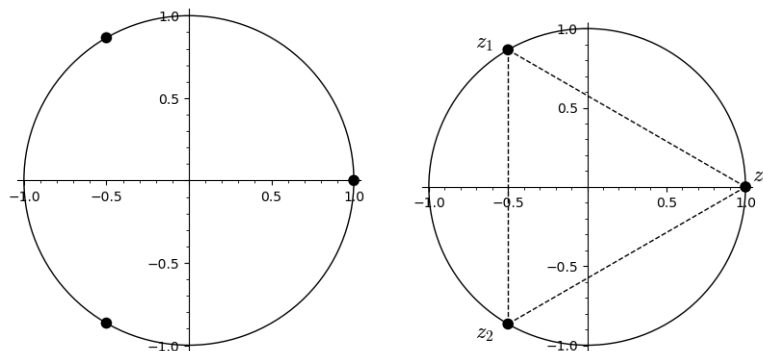
**Solution.** (a) We are interested in solutions of the equation  $z^3 = 1$ , where  $z \in \mathbb{C}$ . By the fundamental theorem of algebra we expect three solutions over  $\mathbb{C}$ . To find them explicitly, we can use de Moivre's formula which provides an easy way for solving equations of the form  $z^n = z_0$ . Indeed, in polar form the equation  $z^3 = 1$  is expressed as  $r^3(\cos(3\varphi) + i \sin(3\varphi)) = 1(\cos 0 + i \sin 0)$ . Verify by simple computations that this equation is satisfied if and only if  $r = 1$  and  $3\varphi = 0 \pmod{2\pi}$ . This means that the three solutions are of the form  $z_0 = \cos 0 + i \sin 0$ ,  $z_1 = \cos \frac{2\pi}{3} + i \frac{2\pi}{3}$ , and  $z_2 = \cos \frac{4\pi}{3} + i \sin \frac{4\pi}{3}$ , or in algebraic form

$$z_0 = 1, \quad z_1 = -\frac{1}{2} + i \frac{\sqrt{3}}{2}, \quad z_2 = -\frac{1}{2} - i \frac{\sqrt{3}}{2}.$$

An appropriate Sage cell that gives directly all three solutions in algebraic form reads as

```
z=var (" z "); solve ( z**3-1==0 , z )
```

It is easy to verify that  $|z_i| = 1$  for any  $i = 0, 1, 2$ , so each of them lies on the unit circle  $S^1$ , see the figure below. In addition, we see that  $z_1^2 = \bar{z}_1 = z_2$ . It follows that the coordinate points of  $z_0, z_1, z_2 = \bar{z}_1$  form an equilateral triangle with sides of length  $\sqrt{3}$  as we can see in the figure (for instance the points  $z_1, z_2$  are conjugate each other, so their distance is  $2\sqrt{3}/2 = \sqrt{3}$ ).



(b) The first claim follows by the fundamental theorem. Now, as for the case  $n = 3$ , de Moivre's formula provides the argument of the roots. In particular, we see that the argument multiplied by  $n$  has to be a multiple of  $2\pi$ . Moreover, the absolute value has to be one, so the roots are of the form

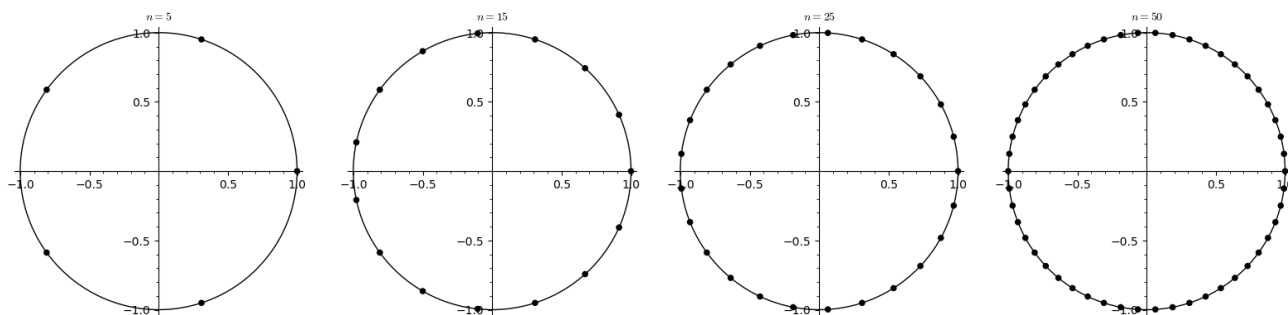
$$z_k = \cos\left(\frac{2k\pi}{n}\right) + i \sin\left(\frac{2k\pi}{n}\right), \quad k = 0, \dots, n-1.$$

Recall now that polygons are geometric objects in the plane composed of points and line segments connected together to close and form a single shape. A regular polygon is a polygon having equal angles and sides of equal length. For the roots derived above we see that the argument of any two consecutive of them, differs by  $2\pi/n$ . Moreover, the absolute value of any root equals to 1. Together these claims prove that the points under question are vertexes of a regular  $n$ -polygon.  $\square$

**1.G.18. Polygons via Sage.** In Sage on can apply the following cell to produce the second figure given above (this already includes the code producing the first simpler figure).

```
n=3;
A=[exp(2*pi*I*k/n) for k in range(n)]
a = plot([],xmin=-1, xmax=1, ymin=-1, ymax=1, figsize=(4,4), rgbcolor=(1/4,1/8,3/4))
a+=list_plot(A, color="black", size=70, xmin=-1,xmax=1)
a+= circle ((0.0 ,0.0) ,1, color="black")
a+=line([(-1/2, (sqrt(3)/2)), (-1/2,-(sqrt(3)/2))], color="black", linestyle="--")
a+=line([(-1/2, (sqrt(3)/2)), (1, 0)], color="black", linestyle="--")
a+=line([(1, 0), (-1/2,-(sqrt(3)/2))], color="black", linestyle="--")
a+= text ("z_2", (-0.65, (sqrt(3)/2)+0.05), color="black", fontsize="16")
a+= text ("z_3", (-0.65, -(sqrt(3)/2) -0.05), color="black", fontsize="16")
a+= text ("z_1", (1.10, 0.07), color="black", fontsize="16")
show(a)
```

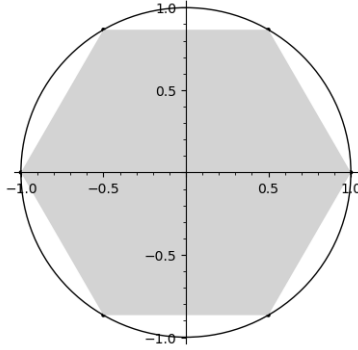
In fact, the commands of the first five lines in the previous code, with only difference the initial  $n$ , can produce similar plots for our favourite  $n$  (though for large  $n$  Sage needs time to answer). Below we present the figures for  $n = 5, 15, 25, 50$ .



Finally it is also possible to produce directly the regular polygons, using the command `polygon2d`. Let us present a Sage cell adapted to the case  $n = 6$ .

```
n=6; A=[exp(2*pi*I*k/n) for k in range(n)]
a = plot([],xmin=-1, xmax=1, ymin=-1, ymax=1, figsize=(4,4))
a+=list_plot(A, color="black", size=10, xmin=-1,xmax=1)
a+=circle((0.0,0.0),1,color="black")
L = [[cos(pi*i/3), sin(pi*i/3)] for i in range(6)]
b=polygon2d(L, color="lightgrey")
show(a+b)
```

This produces the hexagon:



**1.G.19.** Enumerate all ordered pairs  $(x, y) \in \mathbb{R} \times \mathbb{R}$  satisfying the relation  $(x + iy)^{2024} = x - iy$ .

**Solution.** Setting  $z = x + iy$  the given relation becomes  $z^{2024} = \bar{z}$ . Recall that  $|z^2| = |z|^2$  and by mathematical induction (or by de Moivre's formula) one can prove that  $|z^n| = |z|^n$  for any  $n \in \mathbb{Z}$ . Thus, having in mind the relation  $z^{2024} = \bar{z}$  we compute

$$|z|^{2024} = |z^{2024}| = |\bar{z}| = |z| \iff |z|(|z|^{2023} - 1) = 0.$$

From this equation we see that either  $|z| = 0$  or  $|z| = 1$ . The first condition implies that  $(x, y) = (0, 0)$ . If  $|z| = 1$ , then using the relation  $z^{2024} = \bar{z}$  we see that  $z^{2025} = z\bar{z} = |z|^2 = 1$ . However the equation  $z^{2025} = 1$  has 2025 distinct roots. Hence, all together we can enumerate  $2025 + 1$  ordered pairs  $(x, y) \in \mathbb{R} \times \mathbb{R}$  satisfying the given condition.  $\square$

B) MATERIAL ON DIFFERENCE EQUATIONS

**1.G.20. Tilings.** For the Problem 1.B.4 on 2-compositions of natural numbers and for the given solution provide a visual interpretation.

**Solution.** To provide a visual presentation of  $C_n$  (and so also of the Fibonacci numbers) one may think the 1s as squares, and the 2s as dominoes. Then  $C_n$  counts the distinct ways to tile a board of length  $n$  with squares and dominoes, and the result proved in 1.B.4 means that we have  $F_{n+1}$  ways to do so. Below we illustrate the idea for  $n = 4$  and  $n = 5$ . For  $n = 4$  one has the 2-compositions  $1 + 1 + 1 + 1, 1 + 1 + 2, 1 + 2 + 1, 2 + 1 + 1, 2 + 2$ , that is,  $C_4 = 5 = F_5$ . Notice that  $C_4 = C_3 + C_2 = 3 + 2$ .



For  $n = 5$  we can define 8 such decompositions, i.e.,  $C_5 = 8 = F_6$ , as we can see also in the figure below (notice that  $C_5 = C_4 + C_3 = 5 + 3$ ).



$\square$

**1.G.21. Basic properties of Fibonacci numbers.** Show that the Fibonacci numbers  $F_n$  satisfying the following:

- (a) The sum of the first  $n$  Fibonacci numbers equals to  $F_{n+2} - 1$ ;
- (b)  $F_1^2 + F_2^2 + \dots + F_n^2 = F_n F_{n+1}$ ;
- (c)  $F_{m+n} = F_{n-1} F_m + F_n F_{m+1}$ ;
- (d)  $F_{n-1} F_{n+1} - F_n^2 = (-1)^n$ .

**Solution.** (a) We have  $F_1 = F_3 - F_2, F_2 = F_4 - F_3, \dots, F_{n-1} = F_{n+1} - F_n$  and  $F_n = F_{n+2} - F_{n+1}$ . Adding these relations (side by side), we get  $F_1 + F_2 + \dots + F_n = F_{n+2} - F_2$  and the claim follows.

(b) First notice that using the recurrence relation of Fibonacci numbers one may extend these numbers to zero and negative indices, and here we need to know that  $F_0 = F_2 - F_1 = 0$ . Now, since we have

$$F_k F_{k+1} - F_{k-1} F_k = F_k (F_{k+1} - F_{k-1}) = F_k^2,$$

we obtain the relations

$$F_1^2 = F_1 F_2, \quad F_2^2 = F_2 F_3 - F_1 F_2, \quad \dots, \quad F_n^2 = F_n F_{n+1} - F_{n-1} F_n.$$

Adding these relations (side by side) we arrive to the desired formula.

(c)+(d) The last two cases can be proved by induction over  $n$  and we shall prove (d) only, which is known as the *Cassini identity*. For  $n = 1$  it becomes  $F_0 F_2 - F_1^2 = 0 \cdot 1 - 1^2 = -1 = (-1)^1$ , which is true. The induction hypothesis says that the result is true for some arbitrary positive integer  $k$ , which mean that

$$F_{k-1}F_{k+1} - F_k^2 = (-1)^k. \quad (*)$$

Then, for  $n = k + 1$  we compute

$$\begin{aligned} F_k F_{k+2} - F_{k+1}^2 &= (F_{k+1} - F_{k-1})(F_k + F_{k+1}) - F_{k+1}^2 = F_{k+1}F_k + F_{k+1}^2 - F_{k-1}F_k - F_{k-1}F_{k+1} - F_{k+1}^2 \\ &\stackrel{(*)}{=} F_{k+1}F_k - F_{k-1}F_k - (F_k^2 + (-1)^k) = F_{k+1}F_k - F_k(F_{k-1} + F_k) + (-1)^{k+1} \\ &= F_{k+1}F_k - F_k F_{k+1} + (-1)^{k+1} = (-1)^{k+1}, \end{aligned}$$

where we have applied repeatedly the definition of Fibonacci numbers. This means that the claim is true for  $n = k + 1$ , and hence is valid for all  $n \geq 1$ .  $\square$

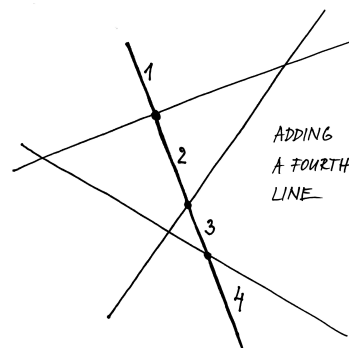


Recurrence relations have a wide range of applications across various fields, often in surprising ways. For example, linear recurrences naturally arise in simple geometric problems, requiring creative problem-solving skills. For your convenience, we have included some examples of these problems below. In Chapter 3 we will cover difference equations again, this time with more advanced applications to expand upon the introduction given in this chapter.

**1.G.22.** Consider  $n$  lines dividing a plane into regions. Find the maximum number of regions that are formed.

**Solution.** Let us denote the number of regions by  $p_n$ . If there is no line in the plane, then the whole plane is one region and  $p_0 = 1$ . If there are  $n$  lines, then adding an additional line will increase the number of regions, exactly by the number of regions that the new line intersects. If no lines are parallel and no three lines intersect at the same point, then the number of regions intersected by the  $(n + 1)$ th line is one more than the number of intersections it has with the existing  $n$  lines. Each intersection divides an existing region into two, resulting in an increase of one region for each intersection. The new line can have at most  $n$  intersections with the existing  $n$  lines. Each segment of the line between two intersections crosses exactly one region, so the new line can cross at most  $n + 1$  regions. Thus we obtain the recurrence relation  $p_{n+1} = p_n + (n + 1)$ , with  $p_0 = 1$ . We can derive an explicit formula for  $p_n$  via Corollary 1.2.3, which we leave to the reader. A more straightforward approach relies on the observation that  $p_n = p_{n-1} + n$ . This gives

$$\begin{aligned} p_n &= p_{n-1} + n = p_{n-2} + (n - 1) + n \\ &= p_{n-3} + (n - 2) + (n - 1) + n = \dots \\ &= p_0 + \sum_{i=1}^n i = 1 + \frac{n(n + 1)}{2} = \frac{n^2 + n + 2}{2}. \end{aligned}$$



Or we can use Sage as we did in 1.B.3:

```
from sympy import Function, rsolve
from sympy.abc import n
p = Function("p")
eq = p(n) - p(n+1) + n + 1
initial = {p(0):1, p(1):2}
rsolve(eq, p(n), initial)
```

$\square$

**1.G.23.** What is the maximum number of areas that a 3-dimensional space can be divided into by  $n$  planes?

**Solution.** Let  $r_n$  be the desired number. Obviously,  $r_0 = 1$ . As in 1.G.22, let us consider  $n$  planes in the space. Now, add an additional plane and determine the maximum number of new regions it can create. The maximum number of new regions is precisely the number of regions that the new plane intersects. What is the maximum number of new regions that can be created?

The number of areas intersected by the  $(n + 1)$ th plane equals the number of regions created by the intersections of this new plane with the  $n$  existing planes. According to the exercise in plane, there can be at most  $1/2 \cdot (n^2 + n + 2)$  such regions. Therefore, we obtain the recurrent formula

$$r_{n+1} = r_n + \frac{n^2 + n + 2}{2}.$$

This equation can be solved directly as follows:

$$\begin{aligned}
 r_n &= r_{n-1} + \frac{(n-1)^2 + (n-1) + 2}{2} = r_{n-1} + \frac{n^2 - n + 2}{2} = r_{n-2} + \frac{(n-1)^2 - (n-1) + 2}{2} + \frac{n^2 - n + 2}{2} \\
 &= r_{n-2} + \frac{n^2}{2} + \frac{(n-1)^2}{2} - \frac{n}{2} - \frac{(n-1)}{2} + 1 + 1 = r_{n-3} + \frac{n^2}{2} + \frac{(n-1)^2}{2} + \frac{(n-3)^2}{2} - \frac{n}{2} - \frac{(n-1)}{2} - \frac{(n-2)}{2} \\
 &\quad + 1 + 1 + 1 \\
 &= \dots = r_0 + \frac{1}{2} \sum_{i=1}^n i^2 - \frac{1}{2} \sum_{i=1}^n i + \sum_{i=1}^n 1 = 1 + \frac{n(n+1)(2n+1)}{12} - \frac{n(n+1)}{4} + n = \frac{n^3 + 6n + 5}{6}.
 \end{aligned}$$

Above we used the formula

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6},$$

which you may want to prove using mathematical induction. □

**1.G.24.** Find the maximum number of areas that the plane can be divided into by  $n$  circles.

**Solution.** A solution to this geometric problem again relies on a recurrence which can be described as follows:

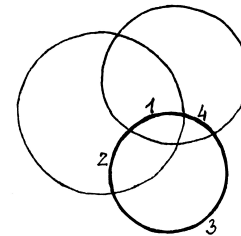
First observe that the  $(n+1)$ th circle intersects the  $n$  existing circles in at most  $2n$  points. For example, see the figure presented at the right-hand side for the case of  $2 + 1$  circles; the third circle intersects the previous two into 4 points (hence  $4 = 2n$  for  $n = 2$ ). Thus, for the maximum number  $p_n$  of areas one can derive the recurrent formula

$$p_{n+1} = p_n + 2n.$$

Clearly  $p_1 = 2$ . Thus for  $p_n$  we obtain

$$\begin{aligned}
 p_n &= p_{n-1} + 2(n-1) = \\
 p_{n-2} + 2(n-2) + 2(n-1) &= \dots = p_1 + \sum_{i=1}^{n-1} 2i = n^2 - n + 2.
 \end{aligned}$$

ADDING THE THIRD CIRCLE



Or we may directly solve the recurrence in Sage, as before:

```

from sympy import Function, rsolve
from sympy.abc import n
p = Function("p")
eq = p(n)-p(n+1)+2*n
initial = {p(1):2, p(2):4}
rsolve(eq, p(n), initial)
    
```

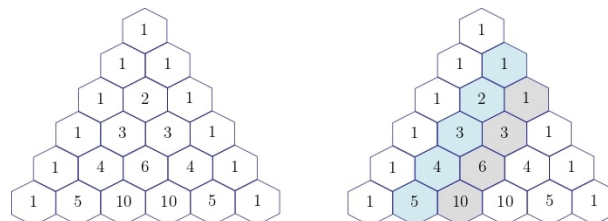
□

**1.G.25.** Determine the maximum number of regions that a 3-dimensional space can be divided into by  $n$  balls. ○

**1.G.26.** Find the number of regions formed when  $n$  distinct planes intersect at a single point in a 3-dimensional space. ○

C) MATERIAL ON COMBINATORICS

**1.G.27. Pascal's triangle via Sage.** Recall that Pascal's triangle is formed by the binomial coefficients, as discussed in the section at the end of the paragraph 1.3.2. In particular, the entry in the  $n$ th row and  $k$ th column of Pascal's triangle is given by the binomial coefficient  $\binom{n}{k}$ . For example for  $n = 6$  we get the following illustration:



Note that in the second figure, the colored "diagonals" represent the natural numbers and triangular numbers, respectively, the latter as illustrated in the first exercise of this section. Present the Pascal's triangle via Sage for  $n = 9$ .

**Solution.** It is sufficient to use the command `binomial(n, i)` and type the cell:

```
[[binomial(n, i) for i in range(n+1)] for n in range(9)]
```

Alternatively, we can give the cell

```
n=8
[[binomial(i, j) for j in range(i + 1)] for i in range(n + 1)]
```

Sage prints out the following:

```
[[1],
 [1, 1],
 [1, 2, 1],
 [1, 3, 3, 1],
 [1, 4, 6, 4, 1],
 [1, 5, 10, 10, 5, 1],
 [1, 6, 15, 20, 15, 6, 1],
 [1, 7, 21, 35, 35, 21, 7, 1],
 [1, 8, 28, 56, 70, 56, 28, 8, 1]]
```

□

**1.G.28.** For any fixed  $n \in \mathbb{N}$ , determine the number of all solutions of the equation  $x_1 + x_2 + \cdots + x_k = n$  for the following two cases:

- (a) over the set  $\mathbb{Z}_{\geq} := \{a \in \mathbb{Z} : a \geq 0\}$  of non-negative integers;  
 (b) over the set  $\mathbb{Z}_{+} := \{a \in \mathbb{Z} : a > 0\}$  of strictly positive integers.

**Solution.** (a) Every solution  $(r_1, \dots, r_k)$  to the equation  $\sum_{i=1}^k x_i = n$  can be uniquely encoded as a sequence of 1s and separators. In particular, the sequence is constructed by writing  $r_1$  1s, followed by a separator, then  $r_2$  1s, followed by another separator, and so on. Consequently, this sequence contains exactly  $n$  1s, and  $k - 1$  separators. Each such sequence uniquely corresponds to a solution of the given equation. Therefore, the number of solutions is equal to the number of such sequences, and the latter is given by  $\binom{n+k-1}{n}$ .

(b) We now look for a solution in the domain of positive integers. We see that the natural numbers  $x_1, \dots, x_k$  provide a solution of the given equation if and only if the non-negative integers  $y_i = x_i - 1, i = 1, \dots, k$  form a solution of the equation

$$y_1 + y_2 + \cdots + y_k = n - k.$$

Using the result obtained in (a), we deduce that there are  $\binom{n-1}{k-1}$  of them. □

**1.G.29. (Trolleybus)** (a) In how many ways can five people seat in a car for five people, under the assumption that only two of them have a driving licence?

(b) In how many ways can 20 passengers and two drivers be seated in a trolleybus for 25 people? ○

**1.G.30. (Flipping coins – I)** We flip a coin six times.

(a) How many distinct sequences of heads and tails are there?

(b) How many sequences with exactly four heads are there?

(c) How many sequences with at least two heads are there? ○

**1.G.31. (Playing with divisions)** Determine the number of all possible ways that the following can happen:

(a) divide 40 identical balls among 4 boys?

(b) divide among three people  $A, B$  and  $C$ , 33 distinct coins such that  $A$  and  $B$  together have twice as many coins as  $C$ .

(c) divide 9 girls and 6 boys into two group such that each group contains at least two boys. ○

**1.G.32.** According to quality, we divide food products into groups  $I, II, III, IV$ . Determine the number of all possible divisions of 9 food products into these groups, such that the numbers of products in groups are all distinct.

**Solution.** If we directly write the considered groups from the elements of  $I, II, III, IV$ , we create combinations of repetitions of the ninth-order from four elements. The number of such combinations is  $\binom{12}{9} = 220$ . □

**1.G.33. (Handshakes)** New players meet in a volleyball team (6 people). How many handshakes are there when everybody shakes once with everybody else? How many handshakes are there if everybody shakes hands once with each opponent after playing a match? ○

**1.G.34.** We need to accommodate 9 people in one four-bed room, one three-bed room and one two-bed room. In how many ways can this be done?

**Solution.** We may assign to the people in the four-bed room the number 1, in the three-bed room number 2 and in the two-bed room number 3. In this way we create permutations with repetitions from the elements 1, 2, 3, where 1 occurs four times, 2 three times and 3 two times. The number of such permutations is  $\frac{9!}{4! \cdot 3! \cdot 2!} = 1260$ .  $\square$

**1.G.35.** In a long-distance race, where the racers start one after another in given time intervals, there were  $k$  racers, among them 3 friends. Determine the number of starting schedules in which no two of the 3 friends start next to each other. For simplicity assume  $k \geq 5$ .

**Solution.** Remaining  $k - 3$  racers can be ordered in  $(k - 3)!$  ways. For the three friends there are then  $k - 2$  places (the start, the end and the  $k - 4$  spaces) where we can put them in  $v(k - 2, 3)$  ways. Using the rule of (combinatorial) product, we obtain

$$(k - 3)! \cdot (k - 2) \cdot (k - 3) \cdot (k - 4) = (k - 2)! \cdot (k - 3) \cdot (k - 4).$$

$\square$

**1.G.36.** There are 32 participants of a tournament. The organisers have stated that the participants must divide arbitrarily into four groups, such that the first one has size 10, the second and the third 8, and the fourth 6. In how many ways can this be done?

**Solution.** We can imagine that from 32 participants we create a row, where first 10 are the first group, next 8 are the second group and so on. On the other hand, there are  $32!$  orderings of all participants. Note that the division into groups is not influenced if we change the order of the people in the same group. Therefore the number of distinct divisions equals to  $\frac{32!}{10! \cdot 8! \cdot 8! \cdot 6!}$ .  $\square$

#### D) MATERIAL ON PROBABILITIES



*Next we will provide supplementary material on basic and conditional probability. Probability theory will be explored in depth in Chapter 10, where we will build on this foundational material by introducing additional key results, such as Bayes' theorem, and discussing its significance in statistics. Additionally, the rich interplay between combinatorics and probability will be examined in Chapter 13, which concludes the book.*

**1.G.37.** For the final exams in the 6th class of all the high-schools in a European country, it is known that:

- 25% of the students fails in the exams of mathematics;
- 18% of the students fails in the exams of physics;
- 10% of the students fails both in the exams of mathematics and physics.

Choose randomly a student belonging to this class.

- (a) Find the probability the student has failed at least in one of these two subjects;
- (b) Find the probability the student has failed in the exams of mathematics, but not in the exams of physics.
- (c) If the student has failed in the exams of physics, find the probability he also failed in the exams of mathematics.
- (d) If the student has failed in the exams of mathematics, find the probability he also failed in the exams of physics.

**Solution.** We define two events: the event M, described by “The candidate has failed in the exams of mathematics”, and the event P, described by “The candidate has failed in the exams of physics”.

- (a) We should compute  $P(M \cup P) = P(M) + P(P) - P(M \cap P)$  which gives  $0.25 + 0.18 - 0.1 = 0.33$ .
- (b) In this case we should compute  $P(M \cap P^c)$ , which equals to  $P(M) - P(M \cap P) = 0.25 - 0.1 = 0.15$ .
- (c) We are looking for the probability  $P(M|P)$ . By definition, we have

$$P(M|P) = \frac{P(M \cap P)}{P(P)} = \frac{0.1}{0.18} \approx 0.55.$$

- (d) In this final case we are looking for the probability  $P(P|M)$ , which is given by

$$P(P|M) = \frac{P(P \cap M)}{P(M)} = \frac{0.1}{0.25} = 0.4.$$

$\square$



**1.G.38.** From ten cards, where exactly one is an ace, we randomly draw a card and put it back. How many times must we do this, so that the probability that the ace is drawn at least once, is greater than 0.9?

**Solution.** Let  $A_i$  be the following event: “at  $i$ th drawing the ace was drawn”. The individual events  $A_i$  are independent, so we know that

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_n)), \quad n \in \mathbb{N}.$$

We are looking for some  $n \in \mathbb{N}$  satisfying

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdots (1 - P(A_n)) > 0.9.$$

Clearly,  $P(A_i) = 1/10$  for all  $i \in \mathbb{N}$ . Therefore, it is sufficient to solve the inequality

$$1 - \left(\frac{9}{10}\right)^n > 0.9,$$

from which one gets

$$n > \frac{\log_a 0.1}{\log_a 0.9}, \quad \text{where } a > 1.$$

Evaluating, we deduce that one must do the drawing at least twenty two times. In particular, a Sage cell verifying this has the following form:

```
x = var("x")
ineq = 1 - 0.9**x > 0.9
sol = solve(ineq, x, algorithm="sympy")
print(sol)
```

Sage prints out  $x > 21.8543453267828$ . □

**1.G.39. (Playing with a dice)**

- (a) Throwing a dice eleven times in a row the result was 4. Determine the probability that the twelfth roll results in 4.
- (b) Throwing  $n$  dice, compute the probability that among the numbers that appeared the values 1, 3 and 6 are not present.
- (c) Throwing a dice determine the conditional probability, that the first die resulted in five under the condition that the sum is 9. Based on this result, decide whether the events “first dice results in five” and “the sum is 9”, are independent. ○

**1.G.40.** We flip a coin five times. For every head, we put a white ball in a hat, for every tail we put in the same hat a black ball. Compute the probability that in the hat there are more black balls than white balls, if there is at least one black ball in the hat.

**Solution.** We can introduce the following two events:

- A – “There are more black than white balls in the hat”;
- H – “There is at least one black ball in the hat”.

Notice the event A has H as a consequence. Denote by  $A^c, H^c$  the complementary events of A and H, respectively. The goal is to compute the conditional probability  $P(A|H)$ . The probability  $P(H^c)$  is  $2^{-5}$ , and the probability of H is the same as the probability  $P(A^c)$  of the complementary event of A. Thus  $P(H) = 1 - 2^{-5}$  and moreover  $P(A) = 1/2$ . Furthermore  $P(A \cap H) = P(A)$ , since the event H contains the event A, as it was mentioned above. This finally gives

$$P(A|H) = \frac{P(A \cap H)}{P(H)} = \frac{\frac{1}{2}}{1 - \left(\frac{1}{2}\right)^5} = \frac{16}{31}.$$
□

**1.G.41.** In a painting exhibition there are 15 paintings and between them 12 are authentic and 3 are copies. A visitor chooses randomly a painting and asks the opinion of an art expert about the authentication of the artwork. The expert can give a correct opinion both about an original painting as much as for a copy, on average 9 out of 10 times. If the expert decides that the painting is authentic, find the probability the expert is right.

**Solution.** (a) Let us consider the events:

- A – “The painting is authentic”;
- B – “The expert considers that the painting is authentic”.

We are looking for the conditional probability  $P(A|B)$ . By the statement we have  $P(A) = \frac{12}{15} = \frac{4}{5}$  and  $P(A^c) = \frac{3}{15} = \frac{1}{5} = 1 - P(A)$ . We also see that the probability of the event B occurring, given that A is true, equals to 9/10, i.e.,  $P(B|A) = \frac{9}{10}$ .

Similarly  $P(B^c|A^c) = \frac{9}{10}$  and moreover  $P(B^c|A) = \frac{1}{10} = P(B|A^c)$ . Thus, by the definition of the conditional probability and the product rule we get

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}. \quad (*)$$

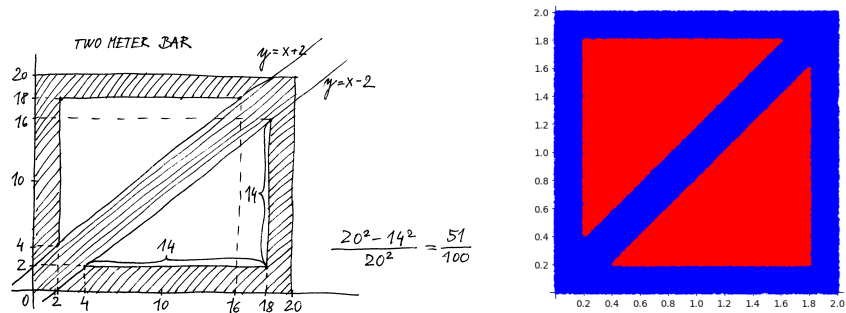
However, by Bayes' formula, see 10.2.7 we have

$$P(B) = P(A) \cdot P(B|A) + P(A^c) \cdot P(B|A^c) = \frac{4}{5} \cdot \frac{9}{10} + \frac{1}{5} \cdot \frac{1}{10} = \frac{37}{50}.$$

Thus, in combination with (\*) we arrive to  $P(A|B) = \frac{36}{37}$ . □

**1.G.42.** A rod of length two meters is randomly divided into three parts. Determine the probability that at least one part is at most 20 cm long.

**Solution.** Random division of a rod into three parts is given by two points of the cut,  $x$  and  $y$  (we first cut the rod in the distance  $x$  from the origin, we do not move it and again cut it in the distance  $y$  from the origin). The sample space is a square  $C$  with side 2 meters, see also the figure below.



If we place the square  $C$  so that its two sides lie on axes in the plane, then the condition that at least one part is at most 20 cm determines in the square a subregion  $O$ , defined by

$$O = \{(x, y) \in C : x \leq 20 \vee x \geq 180 \vee y \leq 20 \vee y \geq 180 \vee |x - y| \leq 20\}.$$

A straightforward computation shows that this subregion has area  $\frac{51}{100}$  times the area of the square. Let us now explain how one can treat this task via Sage. We first use the a cell to generate some random points.

```
num_points = 100_000
#Generate num_points points
x=[RealField().random_element(min=0, max=2) for _ in range(num_points)]
y=[RealField().random_element(min=0, max=2) for _ in range(num_points)]
```

Next we divide the points into two groups corresponding to the case where one piece of rod is less than 20 cm.

```
favourable_events = []
other_events = []
for i, j in zip(x, y):
    if abs(i-j) <= 5:
        favourable_events.append((i, j))
    else:
        other_events.append((i, j))
```

Finally we plot these two sets using red and blue colour (as we see in the figure posed above) via the cell

```
scatter1 = scatter_plot(short_rod_cases, facecolor="blue",
                        marker="o", markersize=10, edgecolor="blue")
scatter2 = scatter_plot(other_cases, facecolor="red",
                        marker="o", markersize=10, edgecolor="red")
show(scatter1+scatter2, axes=true, aspect_ratio=1, ticks=[0.2,0.2])
```

This returns the figure posed above. Recall that we are interested in the ratio of the blue area to the total area of the square. To this end, we can compute its approximate value by typing

```
len(short_rod_cases) / (len(short_rod_cases) + len(other_cases))
```

Sage prints out the estimation 0.51018. □

**1.G.43.** Michael and Alex have a lunch at the school canteen, which operates from 11 to 14. Each of them takes a lunch for 30 minutes, and the arrival time is random. What is the probability that they meet at a given day, if they always sit at the same table? ○

E) MATERIAL ON PLANE GEOMETRY

**1.G.44.** Compute the area  $S$  of a quadrilateral with vertices  $A = [0, -2]$ ,  $B = [1, -1]$ ,  $C = [1, 5]$ , and  $D = [-1, 1]$ . ○

**1.G.45.** Determine the relative position of the lines  $p, q$  in the plane for  $p : 2x - y - 5 = 0$ ,  $q : x + 2y - 5 = 0$ . If they are not parallel, determine the intersection point. ○

**1.G.46.** Determine the sum of the three angles, which are between the vectors  $(1, 1)$ ,  $(2, 1)$  and  $(3, 1)$  respectively and the  $x$ -axis in the plane  $\mathbb{R}^2$ . ○

**1.G.47.** Compute the area of parallelogram with vertices at  $[5, 5]$ ,  $[6, 8]$  at  $[6, 9]$ .

**Solution.** Although such parallelogram is not uniquely determined (the fourth vertex is not given), the triangle with vertices at  $[5, 5]$ ,  $[6, 8]$  and  $[6, 9]$  must be necessarily a half of every parallelogram with these three vertices (one of the sides of the triangle becomes the diagonal of the parallelogram). Therefore the area equals the determinant

$$\begin{vmatrix} 6-5 & 6-5 \\ 8-5 & 9-5 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 3 & 4 \end{vmatrix} = 1 \cdot 4 - 1 \cdot 3 = 1. \quad \square$$

**1.G.48.** Determine the angle  $\varphi$  between the two diagonals  $A_3A_7$  and  $A_5A_{10}$  of a regular dodecagon  $A_0A_1A_2 \dots A_{11}$ .

**Solution.** The angle does not depend neither on the size, nor on the position of the given dodecagon. Consider a dodecagon inscribed in a circle with a diameter of 1, and let  $A_0$  be the point  $[1, 0]$ . The vertices of the dodecagon can be then represented by the twelfth roots of unity in the complex plane. Thus we can write  $A_k = \cos(2k\pi/12) + i \sin(2k\pi/12)$  and hence

$$\begin{aligned} A_3 &= \cos(\pi/2) + i \sin(\pi/2) = i \sim [0, 1], & A_5 &= \cos(5\pi/6) + i \sin(5\pi/6) = -\frac{\sqrt{3}}{2} + \frac{1}{2}i \sim \left[-\frac{\sqrt{3}}{2}, \frac{1}{2}\right], \\ A_7 &= \cos(7\pi/6) + i \sin(7\pi/6) = -\frac{\sqrt{3}}{2} - \frac{1}{2}i \sim \left[-\frac{\sqrt{3}}{2}, -\frac{1}{2}\right], & A_{10} &= \cos(5\pi/3) + i \sin(5\pi/3) = 1/2 - i\frac{\sqrt{3}}{2} \sim \\ & & & \left[1/2, -\frac{\sqrt{3}}{2}\right]. \end{aligned}$$

Combining now the description given in 1.5.7, we deduce that  $\cos \varphi = \frac{1}{2\sqrt{2+\sqrt{3}}}$ . This gives  $\varphi = 75^\circ$ . □

**1.G.49.** Consider the matrices

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad D = \begin{pmatrix} 5 & -b \\ 4 & -4d \end{pmatrix}, \quad E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

with  $a, b, c, d \in \mathbb{R}$ . Solve the matrix equation  $X(E - A^2) + BB^T = D$ . ○

**1.G.50.** Illustrate the expression of a rotation as a composition of shears on an interactive plot, using Sage (use the method from 1.E.18).

**Solution.** A solution goes as follows:

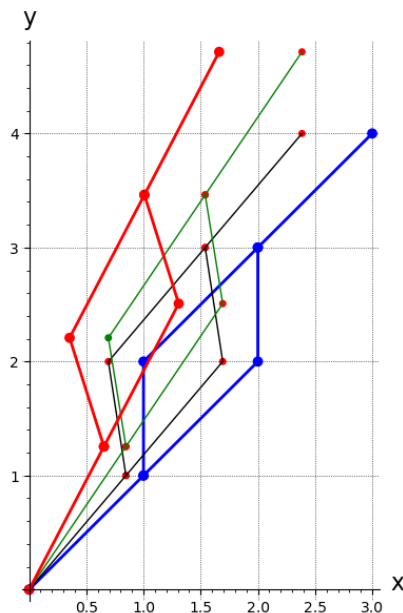
```
# import the the right objects and define the shear functions
from sage.plot.point import point
from sage.plot.line import line
def shear_transformation(x, y, shear_factor):
    new_x = x + shear_factor * y
    new_y = y
    return new_x, new_y
def shear_vert_transformation(x, y, shear_factor):
    new_x = x
    new_y = y + shear_factor * x
    return new_x, new_y
# create an illustrative interactive plot
@interact
def plot_shear(b=(0.0, 1.0, 0.02)):
    # choose points and lines to show
    points1 = [(0, 0), (1, 1), (1, 2), (2, 3)]
    points2 = [(1, 1), (2, 2), (2, 3), (3, 4)]
```

```

# get the shear parameters a and apply shears
a = (sqrt(-b^2 + 1) - 1)/b
sheared_points11 = [shear_transformation(x, y, a) for x, y in points1]
sheared_points12 = [shear_transformation(x, y, a) for x, y in points2]
sheared_points21 = [shear_vert_transformation(x, y, b) for x, y in sheared_points11]
sheared_points22 = [shear_vert_transformation(x, y, b) for x, y in sheared_points12]
sheared_points31 = [shear_transformation(x, y, a) for x, y in sheared_points21]
sheared_points32 = [shear_transformation(x, y, a) for x, y in sheared_points22]
# prepare the objects for the final plot
p1 = point(points1, color='blue', size=50)
p2 = point(points2, color='blue', size=50)
p_sheared11 = point(sheared_points11, color='red', size=25)
p_sheared12 = point(sheared_points12, color='red', size=25)
p_sheared21 = point(sheared_points21, color='green', size=25)
p_sheared22 = point(sheared_points22, color='red', size=25)
p_sheared31 = point(sheared_points31, color='red', size=50)
p_sheared32 = point(sheared_points32, color='red', size=50)
l1 = line(points1, thickness=2)
l2 = line(points2, thickness=2)
l1_sheared1 = line(sheared_points11, thickness=1, color='black')
l2_sheared1 = line(sheared_points12, thickness=1, color='black')
l1_sheared2 = line(sheared_points21, thickness=1, color='green')
l2_sheared2 = line(sheared_points22, thickness=1, color='green')
l1_sheared3 = line(sheared_points31, thickness=2, color='red')
l2_sheared3 = line(sheared_points32, thickness=2, color='red')
# Combine the plots and show the result (choose options)
combined_plot = p1 + p_sheared11 + p_sheared21 + p_sheared31
combined_plot += p2 + p_sheared12 + p_sheared22 + p_sheared32
combined_plot += l1 + l1_sheared1 + l1_sheared2 + l1_sheared3
combined_plot += l2 + l2_sheared1 + l2_sheared2 + l2_sheared3
combined_plot.show(axes_labels=['x', 'y'], gridlines=True, aspect_ratio=1, figsize=8)

```

Executing this code (for one chosen value of  $b = \sin \theta$ , e.g.,  $b = 3$ ) we obtain the following picture:



□

**1.G.51.** (a) Let  $ABC$  be the triangle on the plane with vertices the points  $A = [1, 0]$ ,  $B = [4, -2]$  and  $C = [2, -3]$ , respectively. Reflect  $ABC$  with respect to the  $x$ -axis and illustrate the initial triangle and the result of the reflection via Sage.

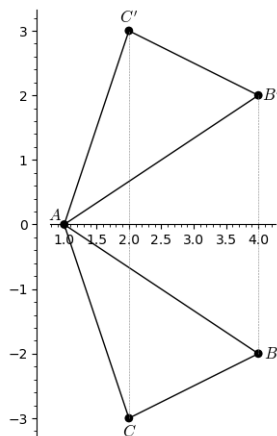
**Solution.** On  $\mathbb{R}^2$  the reflection with respect to the  $x$ -axis is the linear mapping  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  with  $u = (x, y)^T \mapsto f(u) := (x, -y)^T$ . Obviously, the  $x$ -axis is the fixed points set of  $f$ , which means that  $f$  preserves  $A = [1, 0]$ . However, it transforms  $B$  and  $C$  to  $B' = [4, 2]$  and  $C' = [2, 3]$ , respectively, see the figure given below. Alternatively, recall that matrix of  $f$  with respect to the standard basis on  $\mathbb{R}^2$  is given  $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ , such that

$$f(u) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} u = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix}.$$

Viewing  $A, B, C$  as vectors we obtain

$$f(A) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad f(B) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \quad f(C) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 2 \\ -3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

Thus indeed, reflecting the triangle  $ABC$  via  $f$ , we arrive to the triangle  $AB'C'$ , where  $B' = [4, 2]$  and  $C' = [2, 3]$ , respectively. An illustration in Sage can be seen below. Can you reproduce it in your editor? For help, see the cell given in 1.E.26.



□

**1.G.52.** Let  $ABCD$  be the trapezoid on the plane with vertices the points  $A = [0, 0]$ ,  $B = [4, 0]$ ,  $C = [3, 3/2]$  and  $D = [3/2, 3/2]$ . Let  $\{e_1 = (1, 0), e_2 = (0, 1)\}$  be the standard basis, identifying the plane with  $\mathbb{R}^2$ .

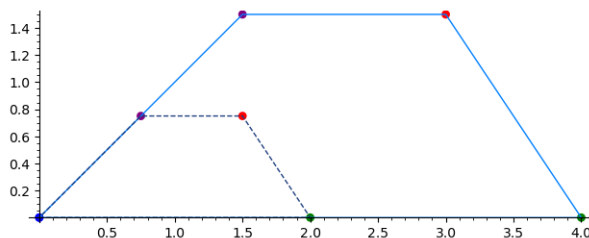
- Transform  $ABCD$  via the homothety  $h_{1/2}$ ;
- Stretch  $ABCD$  in the direction of  $e_1$  with stretching parameter  $c = 3/4$ .
- Stretch  $ABCD$  in the direction of  $e_2$  with stretching parameter  $c = 4/3$ .
- Apply to  $ABCD$  a horizontal shear with parameter  $a = 2$
- Apply to  $ABCD$  a vertical shear with parameter  $a = 1/2$
- Transform  $ABCD$  via the axial symmetry.

Illustrate all the cases.

**Solution.** (a) The matrix of  $h_{1/2}$  is  $\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$ , and we have  $h_{1/2}((x, y)^T) = (x/2, y/2)^T$ . Viewing  $A, B, C, D$  as vectors on the plane, one computes

$$h_{1/2}(A) = [0, 0] = A, \quad h_{1/2}(B) = [2, 0], \quad h_{1/2}(C) = [3/2, 3/4], \quad h_{1/2}(D) = [3/4, 3/4].$$

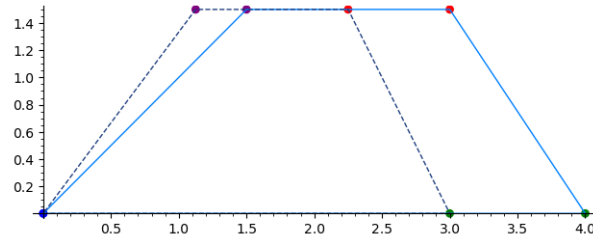
Thus, the new coordinates of the points  $B, C, D$  are  $B' = [2, 0]$ ,  $C' = [3/2, 3/4]$  and  $D' = [3/4, 3/4]$ , respectively, and so  $h_{1/2}$  has shrunk the trapezium (because  $0 < c = 1/2 < 1$ ), as you can see here:



(b) The matrix of stretching in the direction of  $e_1$  only, with stretching parameter  $c = 3/4$  is again diagonal, given by  $\begin{pmatrix} 3/4 & 0 \\ 0 & 1 \end{pmatrix}$ . This means that the corresponding transformation is given by  $(x, y)^T \mapsto (\frac{3x}{4}, y)^T$ , so it preserves the  $y$ -coordinates. Therefore

$$A = [0, 0] \mapsto A, \quad B = [4, 0] \mapsto [3, 0], \quad C = [3, 3/2] \mapsto [9/4, 3/2], \quad D = [3/2, 3/2] \mapsto [9/8, 3/2],$$

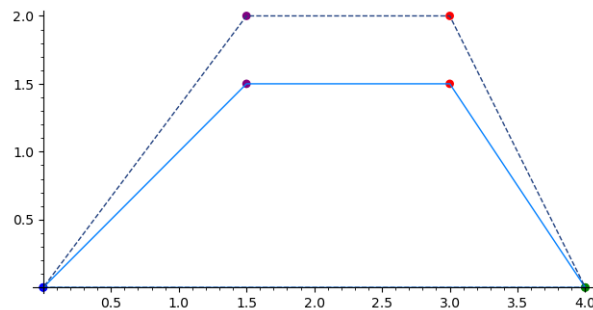
and the initial trapezium  $ABCD$  has been transformed to  $AB'C'D'$ , with  $B' = [3, 0]$ ,  $C' = [9/4, 3/2]$ , and  $D' = [9/8, 3/2]$ , respectively. Hence in this case an illustration looks like as follows:



(c) The stretching here preserves  $x$  and transforms  $y$  by  $\frac{4}{3}$ , i.e., it has the form  $(x, y)^T \mapsto (x, \frac{4y}{3})^T$ . We compute

$$A = [0, 0] \mapsto A, \quad B = [4, 0] \mapsto B, \quad C = [3, 3/2] \mapsto [3, 2], \quad D = [3/2, 3/2] \mapsto [3/2, 2],$$

thus the initial trapezium  $ABCD$  has been transformed to  $ABC'D'$ , with  $C' = [3, 2]$ , and  $D' = [3/2, 2]$ , respectively. As for the illustration we get



(d) As a transformation, the horizontal shear on the plane with parameter  $a = 2$  has matrix  $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ , and thus we get the assignment  $(x, y)^T \mapsto (x + 2y, y)^T$ . We compute

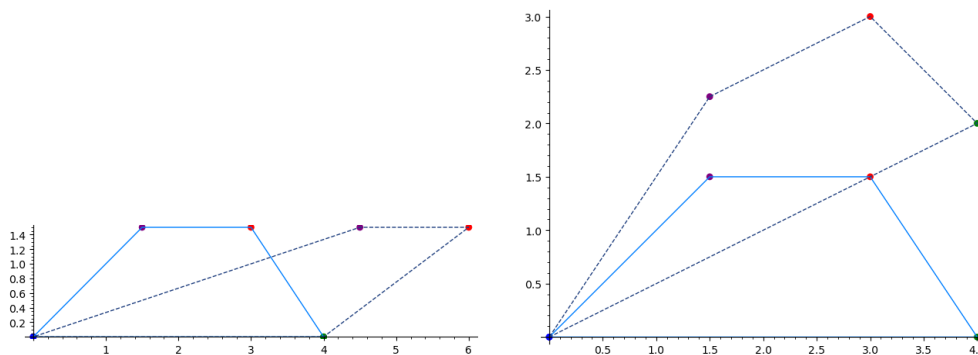
$$A = [0, 0] \mapsto A, \quad B = [4, 0] \mapsto B, \quad C = [3, 3/2] \mapsto [6, 3/2], \quad D = [3/2, 3/2] \mapsto [9/2, 3/2].$$

This means that the new coordinates of initial trapezium are now  $ABC'D'$  with  $C' = [6, 3/2]$  and  $D' = [9/2, 3/2]$ , respectively. Let us for a few postpone its illustration, and present it below together with the illustration of the vertical shear.

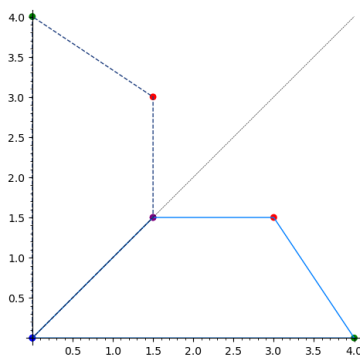
(e) The vertical shear on the plane with parameter  $a = 1/2$  has matrix  $\begin{pmatrix} 1 & 0 \\ 1/2 & 1 \end{pmatrix}$ , so it corresponds to the mapping  $(x, y)^T \mapsto (x, \frac{x}{2} + y)^T$ . We compute

$$A = [0, 0] \mapsto A, \quad B = [4, 0] \mapsto [4, 2], \quad C = [3, 3/2] \mapsto [3, 3], \quad D = [3/2, 3/2] \mapsto [3/2, 9/4],$$

hence in this case the new coordinates of  $ABCD$  are  $AB'C'D'$  with  $B' = [4, 2]$ ,  $C' = [3, 3]$  and  $D' = [3/2, 9/4]$ , respectively. Below, at the l.h.s is presented an illustration of the horizontal shear, while the vertical shear corresponds to the figure at the r.h.s.



(f) Recall that the axial symmetry is the linear transformation  $(x, y)^T \mapsto (y, x)^T$ . Let us present only the illustration of its action on  $ABCD$ , and leave to the reader the mathematical explanation of the figure. Can you predict the vertices which remain fixed (without an illustration or computation).



□

**1.G.53.** Let  $ABCD$  be the rhombus on the plane with vertices the points  $A = [0, 0]$ ,  $B = [5, 0]$ ,  $C = [8.4]$ , and  $D = [3, 4]$ , respectively.

- Rotate  $ABCD$  counter-clockwise around the origin for  $\theta_1 = \pi/3$ .
- Then rotate the rhombus obtained in (a), counter-clockwise around the origin for  $\theta_2 = \pi/3$ .
- Finally rotate the rhombus obtained in (b), counter-clockwise around the origin for another  $\theta_3 = \pi/3$ .
- Determine directly the final position of  $ABCD$ , after imposing the rotations described in (a), (b) and (c).

**Solution.** First observe that the given polygon  $ABCD$  is indeed a rhombus. This is because the vectors  $u_1 = \overrightarrow{AB}$  and  $u_2 = \overrightarrow{AD}$  are of the same length (with respect to the dot product on  $\mathbb{R}^2$ ),  $\|u_1\| = \|u_2\| = 5$ , and moreover the opposite sides are parallel each other. The angle between  $u_1, u_2$  equals to  $\arccos(3/5) \approx 53.13^\circ$ .

For a moment think why after applying a rotation on  $ABCD$ , the new polygon will be again a rhombus. This is because rotations preserves lengths and angles (we will essentially prove this in Chapter 2, studying orthogonal transformations). Thus, rotating  $ABCD$  around the origin, for any possible angle, we will obtain a rhombus with the same characteristics (same lengths and angles as those described for  $ABCD$ ).

The rotation matrix for  $\theta = \pi/3$  is given by  $\begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix}$ , and hence as a linear transformation on  $\mathbb{R}^2$ , the rotation  $R_{\pi/3}$  has the form

$$(x, y)^T \mapsto \left( \frac{x - \sqrt{3}y}{2}, \frac{\sqrt{3}x + y}{2} \right)^T.$$

Based on direct computations, one can now verify the following:

- (a) Under the action of  $R_{\pi/3}$  the points  $A, B, C, D$  (seen as vectors on  $\mathbb{R}^2$ ) are mapped to:

$$A \mapsto A, \quad B \mapsto [5/2, 5\sqrt{3}2], \quad C \mapsto [4 - 2\sqrt{3}, 4\sqrt{3} + 2], \quad D \mapsto \left[ \frac{3}{2} - 2\sqrt{3}, \frac{3\sqrt{3}}{2} + 2 \right].$$

Hence, a counter-clockwise rotation around the origin for  $\theta = \pi/3$  transforms the rhombus  $ABCD$  to  $AB_1C_1D_1$ , where  $B_1 = [5/2, 5\sqrt{3}2]$ ,  $C_1 = [4 - 2\sqrt{3}, 4\sqrt{3} + 2]$  and  $D_1 = \left[ \frac{3}{2} - 2\sqrt{3}, \frac{3\sqrt{3}}{2} + 2 \right]$ , respectively (see the figure below, where we have included all the rotations together).

- (b) Next we need to apply  $R_{\pi/3}$  to  $AB_1C_1D_1$ . We compute

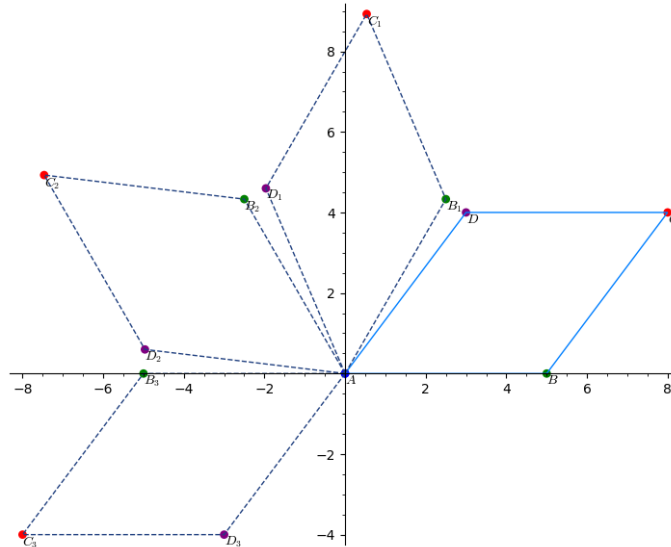
$$A \mapsto A, \quad B_1 \mapsto [-5/2, 5\sqrt{3}2], \quad C_1 \mapsto [-4 - 2\sqrt{3}, 4\sqrt{3} - 2], \quad D_1 \mapsto \left[ -\frac{3}{2} - 2\sqrt{3}, \frac{3\sqrt{3}}{2} - 2 \right].$$

This means that  $AB_1C_1D_1$  has been transformed to the rhombus  $AB_2C_2D_2$ , where  $B_2 = [-5/2, 5\sqrt{3}2]$ ,  $C_2 = [-4 - 2\sqrt{3}, 4\sqrt{3} - 2]$  and  $D_2 = \left[ -\frac{3}{2} - 2\sqrt{3}, \frac{3\sqrt{3}}{2} - 2 \right]$ , respectively.

- (c) Let us finally apply the third rotation, using this time the rhombus  $AB_2C_2D_2$  obtained in (b). We compute

$$A \mapsto A, \quad B_2 \mapsto [-5, 0], \quad C_2 \mapsto [-8, -4], \quad D_2 \mapsto [-3, -4],$$

Thus a counter-clockwise rotation around the origin by  $\pi/3$  transforms the rhombus  $AB_2C_2D_2$  to  $AB_3C_3D_3$ , where  $B_3 = [-5, 0] = -B$ ,  $C_3 = [-8, -4] = -C$  and  $D_3 = [-3, -4] = -D$ , respectively. Let us illustrate all these rotations together in one figure (via Sage), as follows:



The code behind this figure is based on a bit more advanced options from 2D-graphics in Sage, comparing with what we have seen so far (as assigning automatically names to given points in a list). For the interested reader we pose it here:

```
pts = [(0, 0), (5, 0), (8, 4), (3, 4), (5/2, 5*sqrt(3)/2),
       (4-2*sqrt(3), 2+4*sqrt(3)), (3/2-2*sqrt(3), (3/2)*sqrt(3)+2),
       (-5/2, 5*sqrt(3)/2), (-4-2*sqrt(3), 4*sqrt(3)-2), (-3/2-2*sqrt(3), (3/2)*sqrt(3)-2),
       (-5, 0), (-8, -4), (-3, -4)]
pt_names = ['$A$', '$B$', '$C$', '$D$', '$B_1$', '$C_1$', '$D_1$', '$B_2$', '$C_2$',
            '$D_2$', '$B_3$', '$C_3$', '$D_3$']
pt_opt = {'color': 'black', 'horizontal_alignment': 'left', 'vertical_alignment': 'top'}
plt=point2d(pts, color='blue', size=20)
plt += sum(text(name, vector(pt), **pt_opt) for name, pt in zip(pt_names, pts))
plt +=point([0, 0], size=40, color="blue", aspect_ratio=1)
plt +=point([5, 0], size=40, color="green")
plt +=point([8, 4], size=40, color="red")
plt +=point([3, 4], size=40, color="purple")
plt +=polygon([(0,0), (5,0), (8,4), (3, 4)], fill=False,
              rgbcolor= (0, 1/2, 1))#the rhombus ABCD
plt +=point([5/2, 5*sqrt(3)/2], size=40, color="green")
plt +=point([4-2*sqrt(3), 2+4*sqrt(3)], size=40, color="red")
plt +=point([3/2-2*sqrt(3), (3/2)*sqrt(3)+2], size=40, color="purple")
plt +=polygon([(0,0), (5/2, 5*sqrt(3)/2), (4-2*sqrt(3), 2+4*sqrt(3)), (3/2-2*sqrt(3),
              (3/2)*sqrt(3)+2)], fill=False, rgbcolor=(1/8,1/4,1/2), linestyle="--")#the rhombus AB_1C_1D_1
plt +=point([0, 0], size=40, color="blue", aspect_ratio=1)
plt +=point([-5/2, 5*sqrt(3)/2], size=40, color="green")
plt +=point([-4-2*sqrt(3), 4*sqrt(3)-2], size=40, color="red")
plt +=point([-3/2-2*sqrt(3), (3/2)*sqrt(3)-2], size=40, color="purple")
plt +=polygon([(0,0), (-5/2, 5*sqrt(3)/2), (-4-2*sqrt(3), 4*sqrt(3)-2), (-3/2-2*sqrt(3),
              (3/2)*sqrt(3)-2)], fill=False, rgbcolor=(1/8,1/4,1/2), linestyle="--")#the rhombus AB_2C_2D_2
plt +=point([-5, 0], size=40, color="green")
plt +=point([-8, -4], size=40, color="red")
plt +=point([-3, -4], size=40, color="purple")
plt += polygon([(0,0), (-5, 0), (-8, -4), (-3, -4)], fill=False,
              rgbcolor=(1/8,1/4,1/2), linestyle="--")#the rhombus AB_3C_3D_3
plt.show(aspect_ratio=1, figsize=8)
```

(d) This task relies on the principle that composing two rotations results in another rotation, which can be easily determined. In other words, applying a rotation by an angle  $\theta_1$  followed by a rotation by an angle  $\theta_2$  is equivalent to a single rotation by the angle  $\theta_1 + \theta_2$ . In terms of matrices, this implies that

$$\begin{pmatrix} \cos \theta_1 & -\sin \theta_1 \\ \sin \theta_1 & \cos \theta_1 \end{pmatrix} \begin{pmatrix} \cos \theta_2 & -\sin \theta_2 \\ \sin \theta_2 & \cos \theta_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta_1 + \theta_2) & -\sin(\theta_1 + \theta_2) \\ \sin(\theta_1 + \theta_2) & \cos(\theta_1 + \theta_2) \end{pmatrix},$$



that is,  $R_{\theta_1}R_{\theta_2} = R_{\theta_1+\theta_2}$ . To prove this identity, you can use the trigonometric identities provided in part (d) of Problem 1.G.15. For our case we compute

$$R_{\theta_1}R_{\theta_2} = R_{2\pi/3}, \quad \text{and} \quad (R_{\theta_1}R_{\theta_2})R_{\theta_3} = R_{2\pi/3}R_{\theta_3} = R_{2\pi/3}R_{\pi/3} = R_{\pi}.$$

To determine the final position of the initial rhombus  $ABCD$  it is now sufficient to apply a rotation by  $\pi$ . This results in the linear transformation

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \cos \pi & -\sin \pi \\ \sin \pi & \cos \pi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ -y \end{pmatrix} = - \begin{pmatrix} x \\ y \end{pmatrix}.$$

Hence  $ABCD$  will be finally transformed to  $AB'C'D'$ , where  $B' = -B$ ,  $C' = -C$  and  $D' = -D$ , respectively. As a verification, we see that these are the points obtained in step (c), that is  $B' = B_3$ ,  $C' = C_3$  and  $D' = D_3$ , respectively.  $\square$

F) MATERIAL ON RELATIONS AND MAPPINGS

**1.G.54.** Consider a relation  $R$  on the set  $\{1, 2, 3, 4\}$  where  $(a, b) \in R$  if and only if  $a - b$  is divisible by 2. Show that  $R$  is an equivalence relation.

**Solution.** To verify reflexivity we need to show that  $(a, a) \in R$  for all  $a \in \{1, 2, 3, 4\}$ . This is obvious since 0 is divisible by 2. Therefore,  $R$  is reflexive. To verify symmetry we need to show that if  $(a, b) \in R$  then  $(b, a) \in R$ . If  $(a, b) \in R$  the  $a - b$  is divisible by 2, and hence also  $b - a = -(a - b)$  is divisible by 2. Thus  $(b, a) \in R$  and  $R$  is symmetric. To verify transitivity we need to show that if  $(a, b) \in R$  and  $(b, c) \in R$  then  $(a, c) \in R$ . If  $(a, b) \in R$  and  $(b, c) \in R$  then both  $a - b$  and  $b - c$  are divisible by 2 and their sum  $(a - b) + (b - c) = a - c$  is also divisible by 2. Thus  $(a, c) \in R$  and  $R$  is transitive. For example  $(1, 3) \in R$  since  $1 - 3 = -2$  is divisible by 2, and since  $R$  is symmetric we have also  $(3, 1) \in R$ . Then we see that  $(1, 1) \in R$ , as well. Since  $R$  is reflexive, symmetric, and transitive, it is an equivalence relation.  $\square$

**1.G.55.** Determine the number of mappings from the set  $\{1, 2\}$  to the set  $\{a, b, c\}$ . How many of them are surjective and how many are injective?

**Solution.** The element 1 can be mapped to any of  $a, b, c$  and the same for 2. Thus there are exactly  $3^2$  mappings of the set  $\{1, 2\}$  to the set  $\{a, b, c\}$ . None of them can be surjective, because the set  $\{a, b, c\}$  has more elements than the set  $\{1, 2\}$ .

Injectivity requires that the images of 1 and 2 are different. There are three possibilities for the image of 1, and once the image of 1 is given, there remain two possibilities for the image of 2. Thus, the number of injective mappings of the set  $\{1, 2\}$  to the set  $\{a, b, c\}$  is 6.  $\square$

**1.G.56.** Let  $\{a, b, c, d\}$  be a set with a relation  $\{(a, a), (b, b), (a, b), (b, c), (c, b)\}$ . What is the minimal number of elements we have to add to the relation in order to make it an equivalence?

**Solution.** Let us successively ensure the three properties that define an equivalence. First it is the reflexivity. We must add the tuples  $\{(c, c), (d, d)\}$ . Second is the symmetry – we must add  $(b, a)$  and for the third step we must do the so-called transitive closure. Since  $a$  is in relation with  $b$  and  $b$  is in relation with  $c$ , we must add  $(a, c)$  and  $(c, a)$ .  $\square$

**1.G.57.** Determine how many distinct binary relations can be defined between a set  $X$  and the set of all subsets of  $X$ , if the set  $X$  has exactly 3 elements.

**Solution.** First, notice that the set of all subsets of  $X$  has exactly  $2^3 = 8$  elements, and thus the Cartesian product with  $X$  has  $8 \cdot 3 = 24$  elements. Possible binary relations correspond to subsets of this Cartesian product, and hence the total numbers of such relations is  $2^{24} = 16777216$ .  $\square$

**1.G.58.** Find the number of surjective mappings  $f$  from the set  $\{1, 2, 3, 4, 5\}$  to the set  $\{1, 2, 3\}$  such that  $f(1) = f(2)$ .

**Solution.** Every such mappings is uniquely given by the images of the elements  $\{1, 3, 4, 5\}$ , there are exactly that many mappings as there are surjective mappings of the set  $\{1, 3, 4, 5\}$  to the set  $\{1, 2, 3\}$ , that is, 36.  $\square$

**1.G.59.** Determine the number of surjective mappings of the set  $\{1, 2, 3, 4\}$  to the set  $\{1, 2, 3\}$ .  $\circ$

**1.G.60.** List all the relations on the two-element set  $\{1, 2\}$  that are symmetric, but neither reflexive nor transitive.

**Solution.** The reflexive relations are exactly those which contain both pairs  $(1, 1)$ ,  $(2, 2)$ . Hence this excludes the following relations

$$\{(1, 1), (2, 2)\}, \quad \{(1, 1), (2, 2), (1, 2)\}, \quad \{(1, 1), (2, 2), (2, 1)\}, \quad \{(1, 1), (2, 2), (1, 2), (2, 1)\}.$$

We claim that the remaining relations, which are symmetric but not transitive, must contain  $(1, 2)$ ,  $(2, 1)$ . If such a relation contains one of these two (ordered) pairs, by symmetry it must contain also the other. If it contains neither of these pairs, then it is clearly transitive. Thus, from the total number of 16 relations over a two-element set we select

$$\{(1, 2), (2, 1)\}, \quad \{(1, 2), (2, 1), (1, 1)\}, \quad \{(1, 2), (2, 1), (2, 2)\}.$$

It is now clear that each of these 3 relations is symmetric, but neither reflexive nor transitive.  $\square$

**1.G.61.** Determine the number of relations over the set  $\{1, 2, 3, 4\}$ , which are both symmetric and transitive.

**Solution.** Relations of the given properties is an equivalence over some subset of the set  $\{1, 2, 3, 4\}$ . In total,

$$1 + 4 \cdot 1 + \binom{4}{2} \cdot 2 + \binom{4}{3} \cdot 5 + 15 = 52,$$

Thus, there are 52 relations on the set  $\{1, 2, 3, 4\}$  that satisfy the required properties. □

**1.G.62.** Determine all the elements in  $S \circ R$ , if

$$R = \{(2, 4), (4, 4), (4, 5)\} \subset \mathbb{N} \times \mathbb{N}, \quad \text{and} \quad S = \{(3, 1), (3, 2), (3, 5), (4, 1), (4, 4)\} \subset \mathbb{N} \times \mathbb{N}.$$

**Solution.** Consider all choices of two ordered tuples

$$(2, 4), (4, 1), \quad (2, 4), (4, 4), \quad (4, 4), (4, 1), \quad (4, 4), (4, 4)$$

such that the second element of the first ordered tuple (which is a member of  $R$ ) is equal to the first element of the second ordered tuple (which is a member of  $S$ ). Then we obtain  $S \circ R = \{(2, 1), (2, 4), (4, 1), (4, 4)\}$ . This task can also be solved using Sage. First, define a function to compose two given relations as follows

```
def RelCompose(Rel1, Rel2):
    RS = set()
    for (a,b) in Rel1:
        for (c,d) in Rel2:
            if b == c:
                RS.add( (a,d) )
    return RS
```

Now we should introduce  $R, S$ , as follows:

```
R = { (2, 4), (4, 4), (4, 5) }; S = { (3, 1), (3, 2), (3, 5), (4, 1), (4, 4) }
print(RelCompose(R, S))
```

Executing the whole block Sage returns  $\{(4, 4), (2, 4), (4, 1), (2, 1)\}$ . □

**1.G.63.** Let  $R$  be the binary relation between the sets  $A = \mathbb{Z}$  and  $B = \mathbb{R}$ , defined by  $R = \{(0, 4), (-3, 0), (5, \pi), (5, 2), (0, 2)\}$ . Express explicitly  $R^{-1}$  and  $R \circ R^{-1}$ . ○

**1.G.64.** Is there an equivalence relation on the set of all lines in the plane that also serves as an ordering?

**Solution.** An equivalence relation (or ordering relation) must be reflexive, therefore every line must be in relation with itself. Furthermore we require that the relation is both symmetric (equivalence) and antisymmetric (ordering). This implies that each line can only be related to itself. If we define the relation such that two lines are in relation if and only if they are identical, we obtain a very natural relation, which is both an equivalence relation and an ordering. We just need to check that it is transitive, which is easy. Therefore, the only relation that meets the criteria is the identity relation over the set of all lines in the plane. □

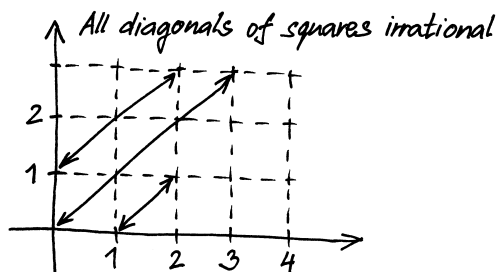
**1.G.65.** We have a set  $\{3, 4, 5, 6, 7\}$ . Write explicitly the following relations and explore their properties:

- i)  $a$  divides  $b$ ,
- ii) Either  $a$  divides  $b$  or  $b$  divides  $a$ ,
- iii)  $a$  and  $b$  have a common divisor greater than one,

○

Solutions to the problems

**1.A.3.** Already the ancient Greeks knew that if we prescribe the area of square as  $a^2 = 2$ , then we cannot find a rational  $a$  to satisfy it. Why? Suppose that we know that  $(p/q)^2 = 2$  for some natural numbers  $p$  and  $q$ , which have no common divisor greater than 1 (otherwise we can further reduce the fraction  $p/q$ ). Then  $p^2 = 2q^2$  is an even number and on the l.h.s  $p^2$  must be even. Therefore so is  $p$ . Now  $p$  is even, thus  $p^2$  must be divisible by 4. But then  $q^2$  is even and so  $q$  must be even too. This certifies that  $p$  and  $q$  both have 2 as a common factor, a contradiction.



**1.A.4.** Let  $p, q$  be two arbitrary rational numbers with  $p \neq q$ . We may assume that  $p < q$ , and similarly is treated the other case. Set

$$\beta := p - \frac{p-q}{\sqrt{2}}.$$

We will show that  $\beta$  is an irrational, that is,  $\beta \in \mathbb{R} \setminus \mathbb{Q}$ , and in addition that  $p < \beta < q$ . The first claim follows because  $1/\sqrt{2}$  is an irrational, in combination with the following two facts:

- The product of a rational with an irrational is an irrational number.
- Adding a rational and an irrational number we obtain an irrational number.

Let us prove the first statement and in a very similar way occurs also the second one. Let  $p = \frac{a}{b} \in \mathbb{Q}$  be a rational number, and let  $x \in \mathbb{R} \setminus \mathbb{Q}$  be an irrational. Then  $xp = px = \frac{xa}{b}$ . Assume that  $xp$  is a rational, that is,

$$xp = \frac{xa}{b} = \frac{r}{s},$$

for some integers  $r, s$ . Multiplying both sides with  $\frac{b}{a}$  we obtain  $x = \frac{rb}{sa}$ . Being the product of the rationals  $\frac{r}{s}$  and  $\frac{b}{a}$ , the number  $x$  should be a rational, a contradiction. Now it remains to prove that  $p < \beta < q$ . We begin with  $0 < \frac{1}{\sqrt{2}} < 1$  and multiply both sides by  $q - p > 0$ . This gives  $0 < \frac{q-p}{\sqrt{2}} < q - p$  and by adding to both sides the number  $p$  we get  $p < p - \frac{p-q}{\sqrt{2}} < q$ . This proves the assertion.

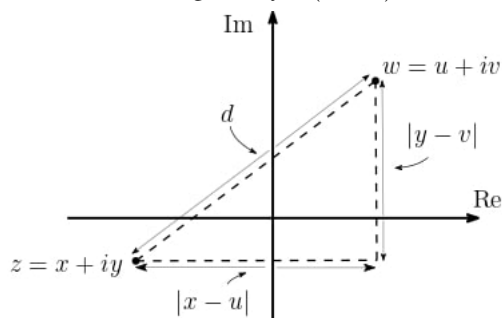
**1.A.7.** By assumption,  $p(x) = 5x^4 + x^2 - 2x + 2$  so the Horner's table has the form

$\rho = 4$	5	0	1	-2	2
		20	80	324	1288
	5	20	81	322	1290

This means that  $p(x) = (x - 4)(5x^3 + 20x^2 + 81x + 322) + 1290$ . In terms of the long division method we have

	$5x^3 + 20x^2 + 81x + 322$				
$x - 4$ )	$5x^4$		$+ x^2$	$- 2x$	$+ 2$
	$- 5x^4 + 20x^3$				
	$20x^3 + x^2$				
	$- 20x^3 + 80x^2$				
	$81x^2 - 2x$				
	$- 81x^2 + 324x$				
	$322x + 2$				
	$- 322x + 1288$				
					1290

**1.A.9.** The horizontal distance between  $z, w$  is the absolute value of the difference of their real parts, that is,  $\text{hor} = |x - u|$ . For an illustration, see the figure below. Note that here one uses the absolute value of  $x - u$ , since depending whether  $z$  lies to the left or to the right of  $w$ , the horizontal distance is given by  $\pm(x - u)$ .



Similarly, the vertical distance between  $z, w$  is the absolute value of the difference of their imaginary parts, that is,  $\text{ver} = |y - v|$ . According to the Pythagorean theorem the distance  $d$  between  $z, w$  satisfies

$$d^2 = \text{hor}^2 + \text{ver}^2 = |x - u|^2 + |y - v|^2$$

and the assertion follows.

**1.A.11.** The distance under question equals to  $\sqrt{13}$ .

**1.A.12.** By assumption,  $i$  is a solution of the given equation. Hence we have that  $i^3 + (1 + i)i^2 + ai + 2 = 0$ . From this equation we compute  $a = 2 + i$ . Thus, we need to solve the equation

$$f(z) = 0, \quad \text{where} \quad f(z) = z^3 + (1 + i)z^2 + (2 + i)z + 2.$$

Since  $f(z)$  is of degree three, over  $\mathbb{C}$  we can specify three roots. In particular, we know that  $i$  is a solution of  $f(z) = 0$ , and we can use this to obtain the corresponding Horner table. Set  $\rho = i$ . We have

$$\begin{array}{r|rrrr} \rho = i & 1 & 1 + i & 2 + i & 2 \\ & & i & i - 2 & -2 \\ \hline & 1 & 1 + 2i & 2i & 0 \end{array}$$

The last entry of the bottom row is 0 since  $i$  is a root of  $f(z) = 0$ . Hence we can write

$$f(z) = z^3 + (1 + i)z^2 + (2 + i)z + 2 = (z - i)(z^2 + (1 + 2i)z + 2i) = (z - \rho)\pi(z)$$

with  $\pi(z) := z^2 + (1 + 2i)z + 2i$ . Now, the discriminant of  $\pi(z)$  is given by

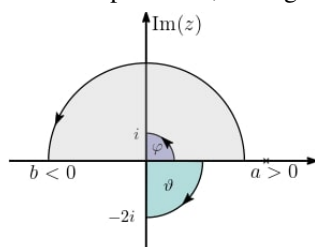
$$\Delta = (1 + 2i)^2 - 4 \cdot 1 \cdot 2i = 1 + 4i - 4 - 8i = -3 - 4i = 1 - 4 - 4i = (1 - 2i)^2.$$

Thus, the rest two roots are given by

$$z_{2,3} = \frac{-(1 + 2i) \pm \sqrt{(1 - 2i)^2}}{2},$$

that is,  $z_2 = -2i$  and  $z_3 = -1$ .

**1.A.16.** (a) This is true, as we can see in figure below. In particular, the argument of a real number is  $k\pi$ , with  $k \in \mathbb{Z}$ .



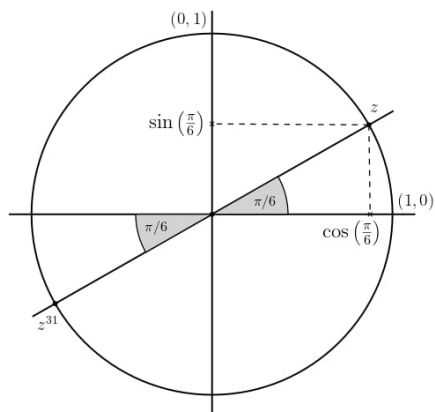
(b) The principal argument of  $i$  equals to  $\varphi = \pi/2$  and is given correctly. However,  $-2i$  has the same principal argument as  $-i$ , i.e.,  $\vartheta = -\pi/2$ , see also the figure. Hence the last statement in (b) is false. One can show that a complex number will be imaginary, if and only if its argument is given by  $\pm\pi/2 + n\pi$ , with  $n \in \mathbb{Z}$ .

**1.A.17.** One can apply the method presented in Problem 1.A.14. This means that for the given  $z \in \mathbb{C}$  we need to compute its magnitude  $r = |z|$  and also its argument. Moreover, a direct calculation shows that

$$|z| = \sqrt{\left(1 + \cos \frac{\pi}{3}\right)^2 + \sin^2 \frac{\pi}{3}} = \sqrt{3}, \quad \cos(\varphi) = \frac{\text{Re}(z)}{|z|} = \frac{1 + \frac{1}{2}}{\sqrt{3}} = \frac{\sqrt{3}}{2}, \quad \sin(\varphi) = \frac{\text{Im}(z)}{|z|} = \frac{1}{2}.$$

Therefore we see that  $\varphi = \pi/6$ . All together, we obtain  $z = \sqrt{3} \left( \cos \frac{\pi}{6} + i \sin \frac{\pi}{6} \right)$ .

**1.A.19.** Recall that the de Moivre's formula states that  $z^n = |z|^n (\cos(n\varphi) + i \sin(n\varphi))$  for any  $n \in \mathbb{Z}$  and  $z \in \mathbb{C}$ , see 1.1.4. We are interested in  $z^{31}$ , where  $z = \cos \frac{\pi}{6} + i \sin \frac{\pi}{6}$  (note that  $z \in S^1 = \{z \in \mathbb{C} : |z| = 1\}$ ). In such tasks it is often useful to draw the diagram corresponding to  $z$ , which for the particular case takes the form



Therefore, we obtain

$$\left( \cos \frac{\pi}{6} + i \sin \frac{\pi}{6} \right)^{31} = \cos \frac{31\pi}{6} + i \sin \frac{31\pi}{6} = \cos \frac{7\pi}{6} + i \sin \frac{7\pi}{6} = -\frac{\sqrt{3}}{2} - i \frac{1}{2}.$$

**1.A.20.** This task highlights the different uses of de Moivre's theorem. Indeed, having in mind the result in 1.1.4, and the identities  $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ ,  $i^2 = -1$ ,  $i^3 = -i$ , we compute

$$\begin{aligned} \cos(3\varphi) + i \sin(3\varphi) &= (\cos(\varphi) + i \sin(\varphi))^3 = \cos^3(\varphi) + 3i \cos^2(\varphi) \sin(\varphi) + 3i^2 \cos(\varphi) \sin^2(\varphi) + i^3 \sin^3(\varphi) \\ &= (\cos^3(\varphi) - 3 \cos(\varphi) \sin^2(\varphi)) + i(3 \cos^2(\varphi) \sin(\varphi) - \sin^3(\varphi)). \end{aligned}$$

A comparison now of the real and the imaginary parts yields the result:

$$\cos(3\varphi) = \cos^3(\varphi) - 3 \cos(\varphi) \sin^2(\varphi) = \cos^3(\varphi) - 3 \cos(\varphi)(1 - \cos^2(\varphi)) = 4 \cos^3(\varphi) - 3 \cos(\varphi),$$

and

$$\sin(3\varphi) = 3 \cos^2(\varphi) \sin(\varphi) - \sin^3(\varphi) = 3(1 - \sin^2(\varphi)) \sin(\varphi) - \sin^3(\varphi) = 3 \sin(\varphi) - 4 \sin^3(\varphi).$$

**1.A.22.** Note that this equation has no rational roots. Substitution into formulas obtained in 1.A.21 yields  $p = b - a^2/3 = -7/3$ ,  $q = -7/27$ . It follows that

$$u = \frac{\sqrt[3]{28 \pm 12\sqrt{-147}}}{6}.$$

We can theoretically choose up to six possibilities for  $u$  (two for the choice of the sign and three independent choices of the cubic root). But we obtain only three distinct values for  $x$ . By substitution into the formulas, one of the roots is of the form

$$\frac{14}{\sqrt[3]{3(28 - 84i\sqrt{3})}} + \frac{\sqrt[3]{28 - 84i\sqrt{3}}}{6} - \frac{1}{3} \doteq 1.247,$$

and similarly for the other two (approximately  $-0.445$  and  $-1.802$ ). Finally observe that even if we have used complex numbers during the computation, all the solutions are real.

**1.B.2.** Set as before  $a = \left(1 + \frac{0.06}{12}\right) = 1.005$  and  $C = 30\,000$ . The condition  $d_k = 0$  gives the equation

$$a^k = \frac{\frac{P}{a-1}}{\frac{P}{a-1} - C} = \frac{200P}{200P - C}.$$

By taking logarithms of both sides, we see arrive to

$$k = \frac{\ln(200P) - \ln(200P - C)}{\ln a}$$

This, for  $P = 500$  gives approximately  $k = 71.5$ . Therefore, Michael will pay for 72 months, with the last repayment being less than €500.

**1.C.3.** (a) The pair of letters b and r can be assumed to be a single indivisible “double-letter”. In total we have six distinct letters and there are  $6!$  words of six indivisible letters. We have to multiply this by two, since the double-letter can be either br, or rb. So the solution is  $2 \cdot 6!$ .

(b) The events in this case form the complement to the part (i), in the set of all rearrangements of the seven-letters. The solution is therefore  $7! - 2 \cdot 6!$ .

**1.C.10.** First, we can place the white tower in any of  $8^2$  positions. Then we have to our disposal  $7^2$  positions (the remaining 7 rows and columns) in which we can place the black tower. Therefore, the total number of ways equals to  $8^2 \cdot 7^2 = 3136$ .

The “inclusion-exclusion” would of course, provide the same result:  $8^2(8^2 - 1) - 8^2 \cdot 7 \cdot 2 = 3136$  (the first terms are all possibilities, the second one the forbidden ones).

**1.D.3.** First divide the favourable cases according to the number of men in the chosen group: 2 or 1. Now, there are eight groups with five people of which one is a man (all women have to be present in such groups, thus it depends only on which man is chosen). Also, there are  $c(8, 2) \cdot c(4, 3) = \binom{8}{2} \cdot \binom{4}{3}$  of groups with two men. This is because we choose two men from eight and then independently three women from four. These two choices can be independently combined and thus using the rule of product we obtain the number of such groups. The total number of groups with five people is  $c(12, 5) = \binom{12}{5}$ . Therefore, the desired probability, being the quotient of the number of favourable outcomes to the total number of outcomes, equals to

$$\frac{8 + \binom{4}{3} \binom{8}{2}}{\binom{12}{5}} = \frac{5}{33}.$$

**1.D.8.** We solve this exercise using the theorem about multiplication of probabilities. This is explained in 1.4.8, based on the conditional probability concept. Here it seems to be obvious:

First we require a red ball, that happens with the probability  $9/16$ . If a red ball was drawn, then in the second round we draw a red ball with the probability  $8/15$  (there are 15 balls in the box, 8 of them are red). Finally, if two red balls were drawn, the probability that a white ball is drawn is  $7/14$  (there are 7 white balls and 7 red balls in the box). Thus we obtain

$$\frac{9}{16} \cdot \frac{8}{15} \cdot \frac{7}{14} = 0.15.$$

**1.E.5.** Eliminate  $t$  to obtain  $q : x - 2y = -5$ . Then solve for  $x$  and  $y$ . The intersection has coordinates  $x = 1, y = 3$ .

**1.E.7.** It is clear that

$$-2 \cdot (-x - \frac{3}{2}y + 2) = 2x + 3y - 4.$$

Thus  $p_1$  and  $p_4$  describe the same line. Moreover, note that  $p_2$  can be rewritten as  $-2x + 2y - 6 = 0$ , thus the lines  $p_2$  and  $p_3$  are parallel and distinct. Also, by eliminating  $t$ , the line  $p_5$  has an equation  $x + y = 0$ , which is not parallel to any other line.

**1.E.12.** We compute

$$A - B = \begin{pmatrix} -2 & 5 \\ -1 & 1 \end{pmatrix}, (A - B)^T = \begin{pmatrix} -2 & -1 \\ 5 & 1 \end{pmatrix}$$

and by matrix multiplication we obtain

$$v = 2 \begin{pmatrix} -2 & -1 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -52 \\ 64 \end{pmatrix}.$$

**1.E.14.** Of course, all these results can be easily computed via Sage and the methods mentioned in the previous subsection. But it is all so easy with 2 by 2 matrices that we could count on fingers, too:

(a) We have  $A - B = \begin{pmatrix} -2 & 5 \\ -1 & 1 \end{pmatrix}$ , so we easily compute:

$$(A - B)^2 = \begin{pmatrix} -2 & 5 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -2 & 5 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -5 \\ 1 & -4 \end{pmatrix}, A^2 = \begin{pmatrix} -10 & 10 \\ -4 & -6 \end{pmatrix}, B^2 = \begin{pmatrix} 4 & 0 \\ -3 & 1 \end{pmatrix}$$

and

$$AB = \begin{pmatrix} -5 & 5 \\ -6 & 2 \end{pmatrix} \neq BA = \begin{pmatrix} 0 & 10 \\ -2 & -3 \end{pmatrix}.$$

Thus, computing  $(A - B)^2$  we may use the formula  $A^2 - AB - BA + B^2$  (but we see that it is false to claim that  $(A - B)^2 = A^2 - 2AB + B^2$ ). Hence in matrix calculus the multiplication is not in general commutative.

(b) We compute  $ABC, BCA,$  and  $CAB$ , which yields:

$$D = 2 \begin{pmatrix} 10 & 35 \\ -4 & 22 \end{pmatrix} - \begin{pmatrix} 8 & 12 \\ -14 & 24 \end{pmatrix} - \begin{pmatrix} 2 & 6 \\ -50 & 30 \end{pmatrix} = \begin{pmatrix} 10 & 52 \\ 56 & -10 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, in general given three arbitrary matrices  $A, B, C$ , we have  $ABC \neq BCA \neq CAB$ .

(c) We compute (in the result, we write  $E = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ ):

$$\frac{c(AB) - (c(AB))^T}{\sqrt{2}} = \frac{c}{\sqrt{2}}(AB - (AB)^T) = \frac{\sqrt{2}c}{2} \left\{ \begin{pmatrix} -5 & 5 \\ -6 & 2 \end{pmatrix} - \begin{pmatrix} -5 & -6 \\ 5 & 2 \end{pmatrix} \right\} = \frac{11\sqrt{2}}{2}cE.$$

(d) This is a direct computational consequence of the formula for determinant. In Chapter 2, we will see that this holds true for square matrices of all sizes.

(f) Let us better handle this one by Sage, where it is convenient to use the cell presented in 1.E.12 and add the code

```
D=det(A)*B-det(B)*A-C.trace()*C; print(det(D)); print(divisors(det(D)))
```

This gives  $-38$ , which is the determinant of  $D$ , and  $[1, 2, 19, 38]$ , which are its divisors.

**1.E.15.** (a) For any two square matrices  $A$  and  $B$  we have

$$(A + B)(A - B) = A^2 - A \cdot B + B \cdot A - B^2.$$

Therefore, the identity  $(A + B) \cdot (A - B) = A^2 - B^2$  is valid if and only if  $AB = BA$ . Thus any pair of matrices which do not commute does the work. A choice is given for instance by

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}.$$

Indeed,  $AB = \begin{pmatrix} 8 & 5 \\ 20 & 13 \end{pmatrix} \neq BA = \begin{pmatrix} 13 & 20 \\ 5 & 8 \end{pmatrix}$ .

(b) Similarly, any two square matrices  $A, B$  satisfy

$$(A + B)(A + B) = A^2 + AB + BA + B^2$$

It follows that  $(A + B)(A + B) = A^2 + 2AB + B^2$  if and only if  $AB = BA$ , as in the first case. Hence the pair of matrices presented above provides an example also for this case.

**1.E.16.** The definition of the linear mappings shows, that knowing the values on  $u = (1, 0)$  and  $v = (0, 1)$  determines the values for all the other vectors:  $(x, y) \mapsto F(xu + yv) = xF(u) + yF(v)$ . This corresponds to placing the vectors  $F(u)$  and  $F(v)$  into the columns of the appropriate matrix  $A$ , and then express the value as  $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto A \begin{pmatrix} x \\ y \end{pmatrix}$ . This answers the question about the role of the columns in the matrix  $A$ , once we have it. Moreover, the affine mappings differ from the linear ones only by having a constant vector added. Thus, an affine mapping is linear if and only if it keeps the origin fixed. In our case, clearly

$$\begin{aligned} F \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) &= \begin{pmatrix} 7 & -3 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix}, \\ G \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) &= \begin{pmatrix} 2 & 2 \\ 4 & -9 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -4 \\ 3 \end{pmatrix} = B \begin{pmatrix} x \\ y \end{pmatrix} + w, \\ F \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad G \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) = \begin{pmatrix} -4 \\ 3 \end{pmatrix}, \end{aligned}$$

and we deduce that the mapping  $F$  is linear, but  $G$  is not.

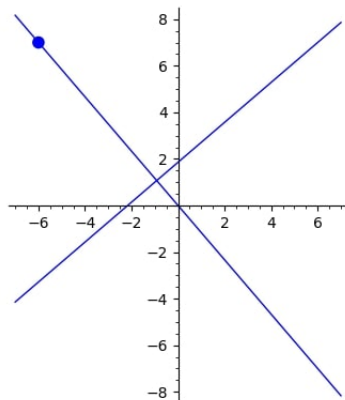
The final claim is obvious from the associativity of matrix multiplication, i.e.,

$$B \left( A \begin{pmatrix} x \\ y \end{pmatrix} \right) + w = (BA) \begin{pmatrix} x \\ y \end{pmatrix} + w, \quad A \left( B \begin{pmatrix} x \\ y \end{pmatrix} + w \right) = (AB) \begin{pmatrix} x \\ y \end{pmatrix} + Aw.$$

Check these computations yourselves and see also the right column in 1.5.4.

**1.E.22.** This is a typical task to be discussed without any explicit coordinates in mind. Let us start with the linear mappings. In the plane, lines are parallel if and only if their directional vectors  $u$  in the parametric representations  $P(t) = P + tu$  are parallel. The images of the lines can be parametrized by  $F(P(t)) = F(P) + tF(u)$ . Now recall that on the plane, two non-zero vectors are parallel if and only if one is a scalar multiple of the other. Thus, the images of parallel lines are clearly parallel as well, except the case, where the image  $F(u) = 0$ . So, strictly speaking, the claim we should prove is false. But we have proved that invertible linear mappings transform parallel lines to parallel lines. Adding translations in order to deal with general affine mappings, does not spoil this property. On the other hand, for example, a shear transformation clearly changes both the angles and the distances, cf. 1.E.17.

**1.E.24.** An ‘‘admissible picture’’ that effectively illustrates the result of the task in 1.E.21, should resemble the figure shown below:



This figure allows you to visually verify the two requirements of the task: first, that the two lines are perpendicular, and second, that the line in question passes through the point  $P = [-6, 7]$ .

To create this plot, you can use a Sage cell with specific options to ensure the correct aspect ratio and figure size. For example, you can type:

```
x=var("x")
y2=-7*x/6; y1=(1/7)*(6*x+13)
fig = plot([], figsize=[4,4])
fig += plot(y1,x, (-7,7), aspect_ratio=1)
fig += plot(y2,x, (-7,7), aspect_ratio=1)
fig += point((-6, 7), size=70)
show(fig)
```

To verify the effect of the different options on the plot, try the following: Modify the Sage cell to use `aspect_ratio="automatic"`. You will observe that the two lines may appear non-perpendicular, which can be misleading. Additionally, changing the `figsize` to `[4, 8]` can produce a distorted or "ugly" figure, further complicating the visual interpretation. By experimenting with these settings in your editor, you can see firsthand how altering the aspect ratio and figure size impacts the clarity and accuracy of your plot.

**1.E.25.** We know that the area equals to the absolute value of the half of the determinant of the matrix, whose first column is given by the vector  $Q - P$  and the second column by the vector  $R - P$ , that is the determinant of the matrix

$$A = \begin{pmatrix} -2 - (-8) & 5 - (-8) \\ 0 - 1 & 9 - 1 \end{pmatrix} = \begin{pmatrix} 6 & 13 \\ -1 & 8 \end{pmatrix}.$$

A simple calculation yields  $\frac{1}{2} |\det(A)| = \frac{1}{2} |(6 \cdot 8 - (13 \cdot (-1)))| = \frac{61}{2}$ . Alternatively, in Sage we can type the cell

```
A=matrix([[6, 13], [-1, 8]]); A.det ( )
```

which returns the determinant of  $A$ . As we will prove in Chapter 2, changing the order of the vectors leads to the change of the sign of the determinant. But the absolute value remains unchanged. Similarly, the transposed matrix (writing the vectors rather in the rows) yields the same determinant. Finally, if the vertices  $P, Q, R$  are ordered in the anti-clockwise direction, the determinant formed by the vectors  $Q - P$  and  $R - P$  is always positive.

If the given coordinates were not in the standard Euclidean coordinate system, we would have to look at their defining frame  $e_1, e_2$ . If their norms are one and they are perpendicular (an orthonormal frame, see the end of 1.5.7), then we do not need to do anything. Otherwise, we should transform the coordinates first into the standard coordinate system, see the methods discussed in 1.E.22 (although area preserving transformation matrices are those with determinant one, e.g., all shears, so this might not be necessary at all – think about it!).

**1.E.26.** Let us denote the quadrilateral  $S$  by  $ABCD$ . Being a polygon, to sketch it in Sage we can apply the command `polygon` or `polygon2d`. Here we use the first command in the following cell:

```
P = polygon([(1,1), (6,1), (11,4), (2,4)], fill=False, color="black")
A =point([1, 1], size=40, color="black")
B =point([6, 1], size=40, color="black")
C =point([11, 4], size=40, color="black")
D =point([2, 4], size=40, color="black")
```

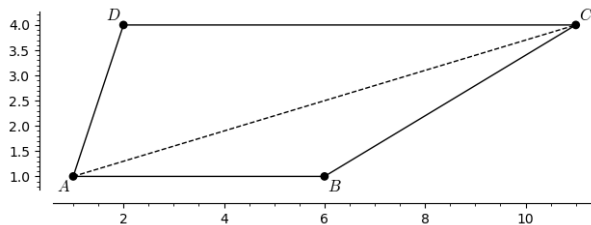


```

A1= text ("A", (0.8, 0.8), color="black", fontsize="12")
B1= text ("B", (6.2, 0.8), color="black", fontsize="12")
C1= text ("C", (11.2, 4.2), color="black", fontsize="12")
D1= text ("D", (1.8, 4.2), color="black", fontsize="12")
l=line([(1, 1), (11, 4)], color="black", linestyle="--")
show(P+A+B+C+D+A1+B1+C1+D1+l)

```

Executing this block we obtain the following illustration:



From this figure we observe that  $S$  is a trapezoid, with bases of length 5 and 9 and a height of 3. Thus, the area of  $S$  is calculated as

$$\text{Area}(S) = \frac{5+9}{2} \cdot 3 = 21.$$

This formula can be derived by noting that shearing transformations do not alter the area of a shape. By selecting an appropriate trapezoid that simplifies the calculations, you can verify this result. Can you identify a suitable trapezoid to use for this purpose?

Alternatively, we can divide  $S$  into two triangles,  $\triangle ABC$  and  $\triangle ACD$ , as shown in the figure. We can then find the area of  $S$  by summing the areas of these two triangles, which can be calculated using the determinants of the corresponding matrices:

$$d_1 = \begin{vmatrix} 6-1 & 11-1 \\ 1-1 & 4-1 \end{vmatrix} = \begin{vmatrix} 5 & 10 \\ 0 & 3 \end{vmatrix}, \quad d_2 = \begin{vmatrix} 11-1 & 2-1 \\ 4-1 & 4-1 \end{vmatrix} = \begin{vmatrix} 10 & 1 \\ 3 & 3 \end{vmatrix},$$

where in the columns are these vectors  $B-A$ ,  $C-A$  (for  $d_1$ ) and  $C-A$ ,  $D-A$  (for  $d_2$ ). In such terms we see that  $\text{Area}(S) = \frac{1}{2}(|d_1| + |d_2|)$ , and this gives 21 as well. Note that the second approach works for all polygonal objects in the plane.

**1.E.28.** The coordinates of the vertex  $C$  can be obtained by rotating the point  $A$  around the centre  $S$  through the angle  $2\pi/3$  in the positive direction. This gives  $C = [\frac{3}{2} - \sqrt{3}, -1 - \frac{\sqrt{3}}{2}]$ . As a verification in Sage you may type

```

a=2*pi/3;x1=0;y1=2;x2=1;y2=0}}
M=matrix([[cos(a), -sin(a)], [sin(a), cos(a)]])
A=vector([x1, y1]); S=vector([x2, y2])
rot=(M*(A-S)+S); show(rot)

```

**1.E.30.** For simplicity, set  $A = [-2, -2]$ ,  $B = [2, -11/6]$ ,  $C = [3, 1]$ ,  $D = [1, 4]$ , and  $E = [-2, 2]$ . The sides  $BC$  and  $CD$  are clearly visible from the position of the point  $[300, 1]$ . On the other hand,  $DE$  and  $EA$  cannot be seen. For the side  $AB$ , we compute

$$\begin{vmatrix} -2-300 & 2-300 \\ -2-1 & -\frac{11}{6}-1 \end{vmatrix} = -302 \cdot (-\frac{17}{6}) - (-298) \cdot (-3) < 0.$$

This implies that the side can be seen from the point  $[300, 1]$ .

**1.E.31.** Order the vertices in the positive direction, that is counter-clockwise:  $[5, 6]$ ,  $[7, 8]$ ,  $[5, 8]$ . Using the corresponding determinants we can determine whether the point  $[0, 1]$  lies to the “left” or to the “right” of the sides of the triangle when we view them as oriented line segments. We have

$$\begin{vmatrix} Q-X \\ R-X \end{vmatrix} = \begin{vmatrix} 7 & 7 \\ 5 & 7 \end{vmatrix} > 0, \quad \begin{vmatrix} R-X \\ P-X \end{vmatrix} = \begin{vmatrix} 5 & 7 \\ 5 & 5 \end{vmatrix} < 0, \quad \begin{vmatrix} P-X \\ Q-X \end{vmatrix} = \begin{vmatrix} 5 & 5 \\ 7 & 7 \end{vmatrix} = 0.$$

We see that the determinants are not all positive, so  $X$  is outside the triangle. In particular, if it is left of some oriented segment (a side of the triangle), the segment is not visible from  $P$ . Because the last determinant is zero, the points  $[0, 1]$ ,  $[5, 6]$  and  $[7, 8]$  lie on a line, and thus the side joining  $P$  and  $Q$  is not visible. The side joining  $Q$  and  $R$  is also not visible, unlike the side joining  $P$  and  $R$ , for which the determinant is negative.

**1.F.4.** we should determine first the three defining properties of an equivalence relation, as functions, and next examine if  $R$  satisfies them. We can do this as follows:

```
def is_reflexive(Rel, A=None):
    if not A:
        A = set(x[0] for x in Rel)
# in this way we define the domain
    return all({(a,a) in Rel for a in A})
def is_symmetric(Rel):
    return all({(b,a) in Rel for (a,b) in Rel})
def is_transitive(Rel):
    return all({(a,d) in Rel for (a,b) in Rel for (c,d) in Rel if b == c})
def relation_summary(Rel):
    A = set(x[0] for x in Rel)
    reflexive = is_reflexive(Rel, A)
    symmetric = is_symmetric(Rel)
    transitive = is_transitive(Rel)
    print(f'Reflexive: {reflexive}')
    print(f'Symmetric: {symmetric}')
    print(f'Transitive: {transitive}')
Rel1 = {('a', 'a'), ('b', 'b'), ('c', 'c'), ('d', 'd'), ('b', 'a'), ('b', 'c'), ('b', 'd')}
relation_summary(Rel1)
```

Sage returns the desired result:

```
Reflexive: True
Symmetric: False
Transitive: True
```

Notice the use of the so called “f-strings” in the print – we are allowed to use expressions whose values should be printed inside the string.

**1.F.5.** From the relationship  $((a, b), (a, b)) \in R$  for all  $a, b \in \mathbb{R}$  it follows that the relation is reflexive. It is also easy to see that the relation is symmetric, since in the equality of the second coordinates we can interchange the left and right side. If  $((a, b), (c, d)) \in R$  and  $((c, d), (e, f)) \in R$ , that is,  $b = d$  and  $d = f$ , then we get that  $((a, b), (e, f)) \in R$ , that is,  $b = f$ . Hence  $R$  is also transitive, and so an equivalence relation. Notice the points in the plane are related if and only if they have the same second coordinate (the line they determine is perpendicular to the  $y$  axis). Thus the corresponding partition divides the plane into the lines parallel with the  $x$  axis.

**1.F.6.** Directly from the definition of the domain and the codomain of a relation we obtain

$$D = \{a, b, c, d, f\} \subset A, \quad I = \{x, y, u, v\} \subset B.$$

To check if a relation is a mapping, we need to ensure that each element in the domain is assigned to exactly one element in the codomain. In our case we deduce that  $R$  is not a mapping since  $(c, x), (c, u) \in R$ , that is,  $c \in D$  has two images.

In Sage, the domain and codomain can be found like this:

```
A = set(["a", "b", "c", "d", "e", "f"])
B = set(["x", "y", "u", "v", "w"])
Rel = set( [ ("a", "v"), ("b", "x"), ("c", "x"), ("c", "u"), ("d", "v"), ("f", "y") ])
# to compute the domain we iterate over R taking the first element of each pair
domain = set(x[0] for x in Rel)
print(f"domain = {domain}")
# to compute the codomain, we take the second element
codomain = set(x[1] for x in Rel)
print(f"codomain = {codomain}")
```

**1.F.9.** In the first case, the map  $f$  is surjective (it is enough to set  $x = 0$ ) but not injective (it is enough to set  $(x, y) = (0, -9)$  and  $(x, y) = (1, 0)$ ). In the second case,  $f$  is an injective mapping (both its coordinates, that is functions  $y = 2x$  and  $y = x^2 + 10$  are clearly increasing over  $\mathbb{N}$ ). The mapping is not surjective (for instance the pair  $(1, 1)$  has no preimage).

**1.G.1.** An easy way is given by Sage, using the command `sum`. Typing in your editor the code

```
# Define symbolic variables
n, k = var("n k")
```

```
# Define the function to sum
f = k
# Compute the symbolic sum
sum1 = sum(f, k, 1, n)
# Show the symbolic sum
show(sum1)
```

Sage's output is  $\frac{1}{2}n^2 + \frac{1}{2}n$ , which represents the evaluated form of the sum  $\sum_{k=1}^n k$ .

However, since the task requires an illustration, we need to explore a different approach. The idea that we will present below likely dates back to the Pythagoreans (6th century BC), but is also closely associated with the German mathematician F. Gauss (1777-1855), one of the most influential scientists of his time. Let  $\Delta_n = 1 + 2 + \dots + (n - 1) + n$  be the sum that we want to compute, and rewrite this with the inverse ordering of the summands, that is,  $\Delta_n = n + (n - 1) + \dots + 2 + 1$ . Adding these two relations, we get

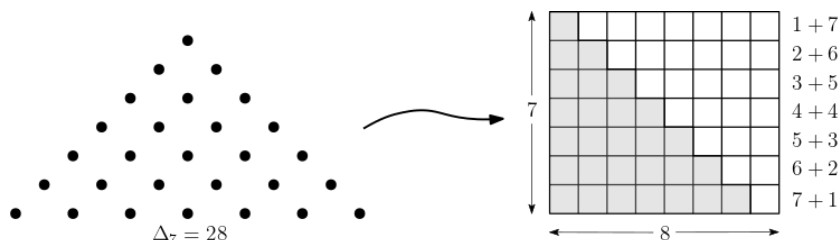
$$2\Delta_n = (n + 1) + (n - 1 + 2) + (n - 2 + 3) + \dots + (2 + (n - 1)) + (1 + n) = (n + 1) + (n + 1) + \dots + (n + 1),$$

where the sum in the r.h.s has  $n$ -factors. Thus  $2\Delta_n = n(n + 1)$ , which completes the proof.

Graphically, the triangular number  $\Delta_n = \frac{1}{2}n(n + 1)$  ( $n = 1, 2, \dots$ ), can be represented by the triangular grid of points (black dots) figured below, where the first row contains a single dot ( $\Delta_1 = 1$ ), and each subsequent row contains one more dot than the previous one (we always move the first dot upwards).



In the figure are presented the first five triangular numbers.<sup>14</sup> To illustrate the simplified proof given above, one converts the above triangle arrangements to half-square arrangements, as in the figure below.



Let us illustrate the case of  $\Delta_7$ . The shaded triangle is the triangle arrangement corresponding to  $\Delta_7$ , while the white triangle occurs after rotating the shaded diagram to obtain the rectangular of size  $7(7 + 1)$ . Hence  $2\Delta_7 = 56 = 7(7 + 1)$ , that is  $\Delta_7 = 28$ . Finally, as in 1.C.7 we see that the number of distinct handshakes or wine glass clinks that can be made among a group of  $n$  people is given by  $\binom{n}{2}$ , and this can be rephrased as  $\Delta_{n-1}$ .<sup>15</sup> This last claim is based on the relation

$$\Delta_n = \binom{n + 1}{2}.$$

As for the two given recurrence relations, they occur very easily and it worths to mention that if we add them, then we return back to the initial definition of  $\Delta_n$ .

**1.G.2.** Suppose that the length of the hill is  $\ell$  km. Then, the bicyclist needs  $\ell/25$  hours to goes up the hill, and  $\ell/75$  hours to goes down the same hill. For instance, if  $\ell = 25$  km, then the cyclist needs one hour to goes uphill and a bit more than half hour for downhill. Since the total distance is  $2\ell$  km, we deduce that the average speed is

$$\frac{2\ell}{\frac{\ell}{25} + \frac{\ell}{75}} = \frac{75 \cdot 2\ell}{4\ell} = \frac{75}{2} = 37.5 \text{ km}.$$

Often, the average speed is called the *harmonic mean* of the two given rates.

**1.G.8.** By assumption,  $P(x) = x^2 + x + 2$ , hence we have

<sup>14</sup>Many mathematicians include 0 as the very first element in the sequence of triangle numbers, but we do not adopt this here.

<sup>15</sup>Each person shakes hands with  $n - 1$  others, but each handshake involves two people, so we count each handshake twice.

$$Q(x) = P(P(x)) - P(x) - 3 = (P(x))^2 + P(x) + 2 - P(x) - 3 = (P(x))^2 - 1 = (P(x) - 1) \cdot (P(x) + 1).$$

This show that  $Q(x)$  is divided by  $P(x) - 1$ , with quotient  $\pi(x) = P(x) + 1 = x^2 + x + 3$ .

**1.G.9.** Suppose that  $P(x) = ax^4 + bx^3 + cx^2 + dx + e$ . Since  $P(0) = 0$  we get  $e = 0$  and so  $P(x) = ax^4 + bx^3 + cx^2 + dx$ . Now, we see that

$$\begin{aligned} P(x+1) - P(x) &= a(x+1)^4 + b(x+1)^3 + c(x+1)^2 + d(x+1) - (ax^4 + bx^3 + cx^2 + dx) \\ &= a(x^4 + 4x^3 + 6x^2 + 4x + 1) + b(x^3 + 3x^2 + 3x + 1) + c(x^2 + 2x + 1) + d(x+1) \\ &\quad - ax^4 - bx^3 - cx^2 - dx \\ &= 4ax^3 + 3(2a+b)x^2 + (4a+3b+2c)x + (a+b+c+d). \end{aligned}$$

Hence, by the relation  $P(x+1) - P(x) = 4x^3$  we get the following system of equations:

$$\{4a = 4, \quad 2a + b = 0, \quad 4a + 3b + 2c = 0, \quad a + b + c + d\}.$$

This has unique solution given by  $a = 1, b = -2, c = 1$  and  $d = 0$ , thus  $P(x) = x^4 - 2x^3 + x^2 = x^2(x^2 - 2x + 1) = x^2(x-1)^2$ , and the claim follows.

**1.G.11.** This task can be challenging, especially at this point. However, we will make every effort to explain the code thoroughly, including our comments within the block, to ensure a clear understanding of each step. We will use the `def` keyword to define a function (or routine), which we name `horner_division`. This function has two inputs: A list of coefficients starting from degree 0, which we may think as  $[a_0, \dots, a_n]$ . The number  $n$  represents the number of coefficients in the polynomial  $p$ . Sage encodes this number by the command `len`. The second entry is the real number  $x_0$ . To perform the division we will combine the `for` loop with the `range()` function. Such combinations are essential in SageMath for handling repetitive tasks, as demonstrated in our example. Recall that the `range()` function generates a sequence of numbers, commonly used in `for` loops. We can customize such sequences using different arguments:

- One argument: `range(stop)` generates numbers from 0 to `stop - 1`.
- Two arguments: `range(start, stop)` generates numbers from `start` to `stop`.
- Three arguments: `range(start, stop, step)` generates numbers from `start` to `stop` with increments of `step`.

With all the necessary tools in place, we can now construct our routine. The code is as follows:

```
def horner_division(p, x0):
    n = len(p) # The number of coefficients in the polynomial
    q = [0] * (n - 1) # Initialize a list to hold the coefficients of the quotient
    r = p[n - 1] # Start with the leading coefficient (highest degree term)
    for i in range(n - 2, -1, -1): # Loop from the second last coefficient to the first
        q[i] = r # Assign the current remainder as a coefficient in the quotient
        r = p[i] + x0 * r # Update the remainder using the current coefficient and x0
    return q, r # Return the list of quotient coefficients and the remainder
```

Our routine is now ready to be tested. Let us use it to prove the second claim.

```
# Example polynomial
# Define the polynomial p(x) = 4 - 3x + 5x^2 - 2x^3 + x^4
p = [4, -3, 5, -2, 1]; x0 = 3
q, r = horner_division(p, x0)
print("Quotient coefficients:", q)
print("Remainder:", r)
```

Sage's output has the form

```
Quotient coefficients: [21, 8, 1, 1]
Remainder: 67
```

Thus dividing  $p(x) = 4 - 3x + 5x^2 - 2x^3 + x^4$  by  $(x - 3)$  we have quotient  $q(x) = 21 + 8x + x^2 + x^3$  and remainder  $r = 67$ .

**1.G.12.** (1) Fractions of the form  $z := \frac{z_1}{z_2}$ , where  $z_1, z_2$  are complex numbers with  $z_2 \neq 0$ , can be written in the form  $z = x + iy$  with  $x, y \in \mathbb{R}$ , after multiplying  $z$  with  $\frac{\bar{z}_2}{z_2}$ . This is because  $z_2 \bar{z}_2 = |z_2|^2 \in \mathbb{R}$ . For the given  $z$  we compute

$$\frac{1 + 2i}{4 - 5i} = \frac{1 + 2i}{4 - 5i} \cdot \frac{4 + 5i}{4 + 5i} = \frac{4 + 5i + 8i + 10i^2}{|4 + 5i|^2} = \frac{-6 + 13i}{4^2 + 5^2} = -\frac{6}{41} + i\frac{13}{41}.$$

This can be quickly verified in Sage by executing the following cell:

```
real((1+2*I)/(4-5*I)) == -6/41; imag((1+2*I)/(4-5*I)) == 13/41
```

For both commands Sage prints out True. You can also obtain the result directly by typing the following command:

```
z=(1+2*I)/(4-5*I); simplify(z)
```

(2) Denote by  $\alpha$  the quantity that we want to compute. We have

$$\alpha := \frac{(1+i)^{2023}}{(1-i)^{2020}} = \left(\frac{1+i}{1-i}\right)^{2020} (1+i)^3$$

and moreover

$$\frac{1+i}{1-i} = \frac{1+i}{1-i} \cdot \frac{1+i}{1+i} = \frac{(1+i)^2}{2} = i.$$

Hence  $\alpha = i^{2020}(1+i)^3 = i^{2020}(2i-2)$ , and it suffices to compute  $i^{2020}$ . A direct computation shows that  $i^4 = 1$  and moreover  $i^8 = 1$ . By induction it follows that  $i^{4n} = 1$  for all  $n \in \mathbb{N}$ . Thus  $i^{2020} = i^{4 \cdot 505} = 1$  and so  $\alpha = 2i - 2$ . A verification by Sage occurs by the cell

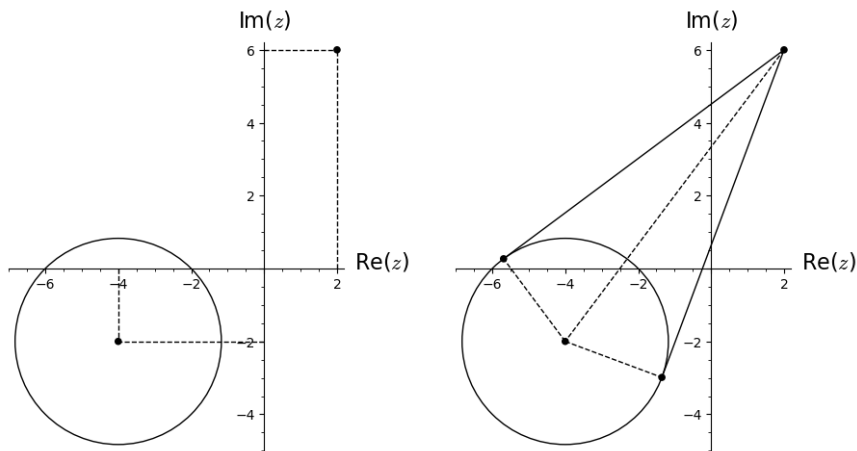
```
a=((1+I)**(2023))/((1-I)**(2020)); simplify(a)
```

(3-4) Both claims are true and indicate that  $S^1$  has the structure of a commutative group under multiplication, as discussed in 1.1.1. Indeed, the first assertion is true because the absolute value function is multiplicative, which implies that  $S^1$  is closed under multiplication. Specifically, for any  $z \in S^1$  we have  $|z| = 1$ , and hence  $z\bar{z} = |z|^2 = 1$ . Moreover, since  $|\frac{1}{z}| = \frac{1}{|z|} = 1$ , we confirm that  $\frac{1}{z} \in S^1$  for any  $z \in S^1$ , ensuring that  $S^1$  is closed under taking reciprocals. Below, we will explore a slightly different but equivalent interpretation of the unit circle, which provides an alternative way to verify these properties.

(5) The statement is true. We will verify this using Sage, and encourage the reader to explore and present alternative methods.

```
z=var("z"); solve(z**3+I==0, z)
```

**1.G.13.** The complex numbers that satisfy the conditions of the task lie on a circle in  $\mathbb{R}^2$ , with centre at  $P_2 = [-4, -2]$  and radius  $\rho$ , where  $\rho$  is the distance between  $z_2$  and  $w$ , i.e.,  $\rho = 2\sqrt{2} = |z_2 - w|$ . Let us illustrate this circle along with the complex numbers  $z_1, z_2$  in a figure (see the figure in the left-hand side).



Now,  $|z_2 - w| = |w - z_2| = |(x+4) + i(y+2)|$ , and we obtain  $\rho^2 = (2\sqrt{2})^2 = 8 = (x+4)^2 + (y+2)^2 = (x^2 + y^2) + 4(2x+y) + 20$ , i.e.,

$$(x^2 + y^2) + 4(2x + y) + 12 = 0. \quad (\dagger)$$

Moreover, if  $b = |z_1 - w| = |(2-x) + i(6-y)|$  is the distance between  $z_1, w$ , then we compute

$$b^2 = (2-x)^2 + (6-y)^2 = (x^2 + y^2) - 4(x+3y) + 40. \quad (\ddagger)$$

Finally, the distance between  $z_1, z_2$  is given by  $d = |z_1 - z_2| = |6 + 8i| = \sqrt{36 + 100} = 10$ . Recall that in Sage this can be verified by the command `abs(6 + 8i)`. We can now proceed with the Pythagorean theorem, which yields the condition  $d^2 = \rho^2 + b^2$ . Due to  $(\ddagger)$  and the above observations, this translates as

$$100 = 8 + (x^2 + y^2) - 4(x+3y) + 40 \iff 52 = (x^2 + y^2) - 4(x+3y).$$

Hence together with  $(\dagger)$  we result to the following system of equations

$$\left\{ \begin{array}{l} (x^2 + y^2) + 4(2x + y) = -12, \\ (x^2 + y^2) - 4(x + 3y) = 52. \end{array} \right.$$

We may multiple the first equation by a minus and then add it to the second one. This gives the equation  $y = -\left(\frac{3}{4}x + 4\right)$ . Replacing now this expression of  $y$  into any of the two equations determining the system (or even in the equation  $(\ddagger)$ ), one arrives to the following quadratic relation that determines  $x$ , namely

$$\frac{25}{16}x^2 + 11x + 12 = 0,$$

with discriminant  $\Delta = 46 > 0$ . Thus, there are two solutions that result in valid triangles, which are illustrated in the figure on the right above. Let us present the explicit form of the solutions:

$$x_{1,2} = \frac{8(-11 \pm \sqrt{46})}{25}, \quad y_{1,2} = -\left(\frac{3}{4}\left(\frac{-88 \pm 8\sqrt{46}}{25}\right) + 4\right).$$

Using Sage one may verify this by typing

```
x, y=var("x, y")
solve([x**2+y**2+4*(2*x+y)+12==0, x**2+y**2-4*(x+3*y)-52==0], x, y)
```

**1.G.25.** Consider first the division of a plane by  $n$  circles. As we know by 1.G.24, the maximum number  $y_n$  of areas a plane can be divided into by  $n$  circles is  $y_n = y_{n-1} + 2(n-1)$ ,  $y_1 = 2$ , that is,  $y_n = n^2 - n + 2$ . For the maximum number  $p_n$  of areas a space can be divided into by  $n$  balls, we obtain the recurrent formula  $p_{n+1} = p_n + y_n$ ,  $p_1 = 2$ , that is,  $p_n = \frac{n}{3}(n^2 - 3n + 8)$ .

**1.G.26.** Let  $x_n$  denote the number of regions in question. When  $n = 1$ , a single plane divides the space into two regions. This means that  $x_1 = 2$ . Now, each new additional plane will intersect all the existing  $n$  planes along lines that divide the new plane into regions. In more details, a new plane will intersect the existing  $n$  planes along  $n$  lines, each of which divides the new plane into additional regions. Thus each new plane adds  $2(n-1)$  new regions. Therefore we derive the following recurrence relation:  $x_n = x_{n-1} + 2(n-1)$ , with the initial condition  $x_1 = 2$ . It is now straightforward to see that the solution has the form  $x_n = n(n-1) + 2$ .

**1.G.29.** (a) For the driver's place we have two choices and the other places are then arbitrary, that is, for the second seat we have four choices, for the third three choices, then two and then 1. That makes  $2 \cdot 4! = 48$  ways.

(b) Similarly in the bus we have two choices for the driver, and then the other driver plus the passengers can be seated among the 24 seats arbitrarily. First choose the seats to be occupied, that is,  $\binom{24}{1}$ . Among these seats the people can be seated in  $21!$  ways. The solution is  $2 \cdot \binom{24}{1} 21! = \frac{24!}{3}$  ways.

**1.G.30.** (a)  $2^6 = 64$ .

(b)  $\binom{6}{4} = 15$ .

(c) No head is one possibility  $\binom{6}{0} = 1$ , one head is  $\binom{6}{1} = 6$ . Thus there are 7 sequences with at most one head and the result is  $64 - 7 = 57$ .

**1.G.31.** (i) Let us add three matches to the 40 balls. If we order the balls and matches in a row, the matches divide the balls in 4 sections. We order the boys at random, give the first boy all the balls from the first section, give the second boy all the balls from the second section and so on. It is now evident that the result is  $\binom{43}{3} = 12\,341$ .

(ii) From the problem statement it is clear that  $C$  must receive 11 coins. That can be done in  $\binom{33}{11}$  ways. Each of the remaining 22 coins can be given either to  $A$  or to  $B$ , which gives  $2^{22}$  ways. Using the rule of product we obtain the result  $\binom{33}{11} \cdot 2^{22}$ . (iii) We divide the boys and the girls independently. Thus the answer is  $2^9(2^5 - 7) = 12800$ .

**1.G.33.** Each pair of players shakes hands at the introduction. As we know the number of handshakes is the combination  $c(6, 2) = \binom{6}{2} = 15$ . After a match each of the six players shakes hands six times (with each of six opponents). Hence the required number is  $6^2 = 36$ .

**1.G.39.** (a) The answer is  $1/6$ .

(b) One can reformulate this task, as we are throwing the dice  $n$  times. The probability that the first roll does not result into 1, 3 or 6 is  $1/2$ . The probability that neither the first nor the second roll is clearly  $1/4$ . This is because the result of the first roll does not influence the result of the second roll. Because the event determined by the result of a given roll and the event determined by the result of another roll are always (stochastically) independent, we deduce that the probability is  $1/2^n$ .

(c) If we denote the event "first dice resulted in five" by  $A$  and the event "the sum is 9" by  $H$ , then it holds

$$P(A|H) = \frac{P(A \cap H)}{P(H)} = \frac{\frac{1}{36}}{\frac{4}{36}} = \frac{1}{4}.$$

Note that the sum 9 occurs when the first die is 3 and the second 6, the first is 4 and the second 6, the first is 5 and the second is 4, or the first is 6 and the second is 3. Of those four results (that have the same probability) only one is favourable to the event  $A$ . Since the probability of  $A$  is clearly  $1/6 \neq 1/4$ , the events are not mutually independent.

**1.G.43.** The space of all possible events can be viewed as a square  $3 \times 3$ . Denote by  $x$  the arrival time of Michael and by  $y$  the arrival time of Alex. In such terms we can claim that these two persons will meet if and only if  $|x - y| \leq \frac{1}{2}$ . This inequality determines in the square of all possible events the area whose volume is  $11/36$  of the volume of the whole square. Thus, this is also the probability of the event.

**1.G.44.** Dividing the quadrilateral into two triangles  $ABC$  and  $ACD$  with areas  $S_1$  and  $S_2$  we obtain

$$S = S_1 + S_2 = \frac{1}{2} \begin{vmatrix} 1-0 & 1-0 \\ -1+2 & 5+2 \end{vmatrix} + \frac{1}{2} \begin{vmatrix} 1-0 & -1-0 \\ 5+2 & 1+2 \end{vmatrix} = \frac{1}{2} (7-1) + \frac{1}{2} (3+7) = 8.$$

**1.G.45.** Eliminating  $y$  yields  $(4x - 2y - 10) + (x + 2y - 5) = 0$ , from which  $x = 3$ , and hence  $y = 1$ . Hence the lines intersect at the point  $P = [3, 1]$ .

**1.G.46.** The given vectors correspond to complex numbers  $1 + i$ ,  $2 + i$  and  $3 + i$ . We are to find the sum of their arguments. According to de Moivre's formula this equals the argument of their product. Their product is  $(1 + i)(2 + i)(3 + i) = (1 + 3i)(3 + i) = 10i$ , which is a purely imaginary number with argument  $\pi/2$ . So the sum we are looking for is  $\pi/2$ .

**1.G.49.** Clearly,  $A^2 = A$  and hence  $E - A^2 = E - A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ . Also,  $B$  is symmetric, i.e.,  $B = B^T$ , and hence

$BB^T = B^2 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ . Thus, the equation  $X(E - A^2) + BB^T = D$  is equivalent to  $\begin{pmatrix} 5 & b+4 \\ 4 & d+5 \end{pmatrix} = \begin{pmatrix} 5 & -b \\ 4 & -4d \end{pmatrix} \iff \begin{pmatrix} 0 & 2b+4 \\ 0 & 5d+5 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ . Its solution is given by  $b = -2$ ,  $d = -1$ , and  $a, c \in \mathbb{R}$  free. In Sage we can directly compute the expression  $X(E - A^2) + BB^T - D$  as follows:

```
var("a, b, c, d")
X=matrix([[a, b], [c, d]]); D=matrix([[5, -b], [4, -4*d]])
A=matrix([[1, 0], [0, 0]]); B=matrix([[1, 2], [2, 1]])
E = identity_matrix(2)
show(X*E-X*A^2+B*(B.transpose())-D)
```

**1.G.59.** The number we need to determine is obtained by subtracting the number of non-surjective mappings from the number of all mappings. The number of all mappings is  $V(3, 4) = 3^4$ . Non-surjective mappings have either a one element, or a two element codomain. There are just three mappings with a one element codomain. The number of mappings with a two-element codomain is  $\binom{3}{2}(2^4 - 2)$  (there are  $\binom{3}{2}$  ways to choose the codomain and for a fixed two-element codomain there are  $2^4 - 2$  ways how to map four elements onto them). Therefore, the number of surjective mappings is  $3^4 - \binom{3}{2}(2^4 - 2) - 3 = 36$ .

**1.G.63.** We see that  $R^{-1} = \{(4, 0), (0, -3), (\pi, 5), (2, 5), (2, 0)\}$ . Moreover,

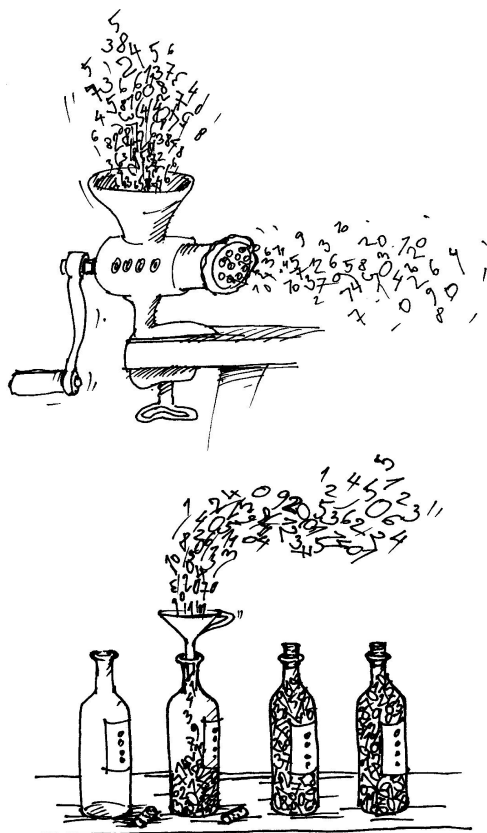
$$R \circ R^{-1} = \{(4, 4), (0, 0), (\pi, \pi), (2, 2), (4, 2), (\pi, 2), (2, \pi), (2, 4)\}.$$

**1.G.65.** Here are the solutions:


- i)  $(3, 3), (4, 4), (5, 5), (6, 6), (7, 7), (3, 6)$ , check that it is an ordering relation.
- ii) again  $(i, i)$  for  $i = 1, \dots, 7$  and additionally  $(3, 6), (6, 3)$ , check that it is an equivalence relation.
- iii)  $(i, i)$  for  $i = 1, \dots, 7$  and also  $(3, 6), (6, 3), (4, 6), (6, 4)$ . Check that it is not an equivalence, since transitivity does not hold.

## Elementary linear algebra

Can't you count with scalars yet?  
 – no worry, let us go straight to matrices...




### A. Vectors and matrices

 The vectors ( $n$ -tuples of scalars) and matrices (2-dimensional arrays of scalars) are the backbone of most of the computational power in practical tasks in all disciplines relying on any kind of data analytics or numerics (e.g., physics, biology, chemistry, engineering, or economics). In this chapter we shall focus on matrix calculus and illustrate some of its simple applications.

To start with, we display a few elementary tasks on matrices and vectors, extending the discussion in the first chapter (see, e.g., 1.E.12 and 1.E.14). Recall that the matrix multiplication works only if the matrix on the left has got the same

In the previous chapter we warmed up by considering relatively simple problems which did not require any sophisticated tools. It was enough to use addition and multiplication of scalars. In this and subsequent chapters we shall add more sophisticated thoughts and concepts.

First we restrict ourselves to concepts and operations consisting of a finite number of multiplications and additions to a finite number of scalars. This will take us three chapters and only then will we move on to infinitesimal concepts and tools. Typically we deal with finite collections of scalars of a given size. We speak about “linear objects” and “linear algebra”. Although it might seem to be a very special tool, we shall see later that even more complicated objects are studied mostly using their “linear approximations”.

 In this chapter we will work with finite sequences of scalars. Such sequences arise in real-world problems whenever we deal with objects described by several parameters, which we shall call coordinates. Do not try much to imagine the space with more than three coordinates. You have to live with the fact that we are able to depict only one, two or three dimensions. However, we will deal with an arbitrary number of dimensions. For example, observing any parameter in a group of 500 students (for instance, their study results), our data will have 500 elements and we would like to work with them. Our goal is to develop tools which will work well even if the number of elements is large.

Do not be afraid of terms like field or ring of scalars  $\mathbb{K}$ . Simply, imagine any specific domain of numbers. Rings of scalars are for instance integers  $\mathbb{Z}$  and all residue classes  $\mathbb{Z}_k$ . Among fields we have seen only  $\mathbb{R}$ ,  $\mathbb{Q}$ ,  $\mathbb{C}$  and residue classes  $\mathbb{Z}_k$  for  $k$  prime.  $\mathbb{Z}_2$  is very specific among them, because the equation  $x = -x$  does not imply  $x = 0$  here, whereas in nearly every other field it does.

### 1. Vectors and matrices

In the first two parts of this chapter, we will work with vectors and matrices in the simple context of finite sequences of scalars. We can imagine working with integers or residue classes as well as real or complex numbers. We hope to illustrate how easily a concise and formal reasoning can lead to strong results valid in a much broader context than just for real numbers.



number of columns as the number of rows of the matrix on the right. The number of rows in the result coincides with that of the left matrix, while its number of columns equals the number of columns of the right-hand matrix.

**2.A.1. Matrix multiplication.** Check the computations:

$$\begin{aligned} \text{(a)} \quad & \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 5 & 4 \end{pmatrix}, \\ \text{(b)} \quad & \begin{pmatrix} 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 7 \end{pmatrix}, \\ \text{(c)} \quad & \begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 2 & 1 \\ 1 & 1 & -2 & -3 \\ 3 & 2 & 1 & 0 \end{pmatrix} \\ & = \begin{pmatrix} 12 & 7 & 1 & -5 \\ 3 & 0 & 5 & 4 \end{pmatrix}, \\ \text{(d)} \quad & (1 \ 3 \ -3) \begin{pmatrix} 1 & -2 & 3 \\ 3 & 2 & 1 \\ 1 & -1 & -4 \end{pmatrix} = (7 \ 7 \ 18), \\ \text{(e)} \quad & (1 \ 2 \ -2) \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix} = (-2) = -2. \end{aligned}$$

**Remark.** As we already mentioned in Chapter 1, the multiplication of matrices is not commutative in general. Parts (a) and (b) in 2.A.1 illustrate this fact for  $2 \times 2$  matrices. On the other hand, part (e) illustrates that multiplying rows and columns produces the so called *scalar products* of vectors (also called *dot products*), and our approach to distances and angles will rely on them. For two vectors like  $u = (2, 1, 3)^T$  and  $v = (1, 2, -2)^T$ , we write  $\langle u, v \rangle$  or  $u \cdot v$  for  $v^T u = -2$ .

As usual, we use the superscript  $T$  to indicate the transposition of the matrices (i.e., we write the rows of  $A$  as columns in  $A^T$ ), and  $E$  for the identity matrix and its dimension will be clear from the context.

**2.A.2.** For the following matrices and vectors

$$A = \begin{pmatrix} 1 & 1 \\ 0 & \sqrt{2} \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 4 \\ 4 & 0 \end{pmatrix}, \quad u = \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix}, \quad v = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

compute the given expressions:

- 1)  $(A - E)(A + E)$ ,
- 2)  $A(Bu)$ ,
- 3)  $(A - B)(A + B)u$ ,
- 4)  $aA^2 + aB^2$  ( $a \in \mathbb{R}$ ),
- 5)  $u \cdot v - 4\sqrt{2}$ ,
- 6)  $4A^3 - B^2$ ,
- 7)  $\sqrt{2}u - \frac{1}{2}v$ ,
- 8)  $\frac{\sqrt{2}}{2}u - ABu + 2v$ ,
- 9)  $A^2u - B^2v$ ,
- 10)  $(B^2u) \cdot (4v)$ .

**2.A.3.** Compute all tasks in 2.A.2 in Sage.

**Solution.** A solution goes as follows:

Later, we follow the general terminology where the notion of vectors is related to fields of scalars only.

**2.1.1. Vectors over scalars.** For now, a *vector* is for us an ordered  $n$ -tuple of scalars from  $\mathbb{K}$ , where the fixed  $n \in \mathbb{N}$  is called *dimension*.

We can add and multiply scalars. We will be able to add vectors, but multiplying a vector will be possible only by a scalar. This corresponds to the idea we have already seen in the plane  $\mathbb{R}^2$ . There, addition is realized as vector composition (as composition of arrows having their direction and size and compared when emanating from the origin). Multiplication by scalar is realized as stretching the vectors.

A vector  $u = (a_1, \dots, a_n)$  is multiplied by a scalar  $c$  by multiplying every element of the  $n$ -tuple  $u$  by  $c$ . Addition is defined coordinate-wise.

BASIC VECTOR OPERATIONS

For all vectors  $u, v$  and scalars  $c$  we define

$$\begin{aligned} u + v &= (a_1, \dots, a_n) + (b_1, \dots, b_n) \\ &= (a_1 + b_1, \dots, a_n + b_n) \\ c \cdot u &= c \cdot (a_1, \dots, a_n) = (c \cdot a_1, \dots, c \cdot a_n). \end{aligned}$$

For vector addition and multiplication by scalars we shall use the same symbols as for scalars, that is, respectively, plus and either dot or juxtaposition, i.e.,  $cu = c(a_1, \dots, a_n) = (ca_1, \dots, ca_n)$ .

**The vector notation convention.** We shall not, unlike many other textbooks, use any special notations for vectors and leave it to the reader to pay attention to the context. For scalars, we shall mostly use letters from the beginning of the alphabet, for the vectors from the end of the alphabet. The middle part of the alphabet can be used for indices of variables or components and also for summation indices.

In the general theory in the end of this chapter and later, we will work exclusively with fields of scalars when talking about vectors. Now we will work with the more relaxed properties of scalars as listed in 1.1.1.

For vector addition in  $\mathbb{K}^n$ , the properties (CG1)–(CG4) (see 1.1.1) clearly hold with the zero element being (notice we define the addition coordinate-wise)  $0 = (0, \dots, 0) \in \mathbb{K}^n$ . We are purposely using the same symbol for both the zero vector element and the zero scalar element. Next, let us notice the following basic properties of vectors:

VECTOR PROPERTIES

For all vectors  $v, w \in \mathbb{K}^n$  and scalars  $a, b \in \mathbb{K}$  we have

- (V1)  $a \cdot (v + w) = a \cdot v + a \cdot w$
- (V2)  $(a + b) \cdot v = a \cdot v + b \cdot v$
- (V3)  $a \cdot (b \cdot v) = (a \cdot b) \cdot v$
- (V4)  $1 \cdot v = v$

```
E=matrix.identity(2)
A=matrix([[1, 1], [0, sqrt(2)]])
B=matrix([[1, 4], [4, 0]])
u=vector([0, sqrt(2)])
v=vector([1, 4])
show((A-E)*(A+E))
show(A*B*u)
show((A-B)*(A+B)*u)
show(A^2+B^2)
u.dot_product(v)
show(4*A^3-B^2)
show(sqrt(2)*u-(1/2)*v)
show((sqrt(2)/2)*u-A*B*u+2*v)
show(A^2*u-B^2*v)
(B^2*u).dot_product(4*v)
```

□

Doing computations with matrices, we always recommend to use Sage to verify your formal computations, especially for large size matrices.

**2.A.4.** Determine explicitly each of the following vectors (by hand), and next verify your answer in Sage:

- 1)  $u_1 := (A - E)^2(4B)u$ ,    3)  $u_3 := A(Bu)$ ,
- 2)  $u_2 := (A - E)^T(B^2u)$ ,    4)  $u_4 := B(Au)$ ,

where  $E$  is the identity  $3 \times 3$  matrix and  $A, B, u$  are respectively given by

$$A = \begin{pmatrix} 1 & \sqrt{2} & \pi \\ 0 & 1 & 1 \\ -\pi & 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & \pi \\ -1 & 0 & 1 \end{pmatrix}, \quad u = \begin{pmatrix} \pi \\ 0 \\ \sqrt{2} \end{pmatrix}.$$

○

*While explicit computations in high dimensions (i.e., dealing with vectors with large number of components) can hardly be imagined without the use of computer aided mathematics software, like Sage, the understanding of structure and properties is often based on visual perception. Of course, we are then limited to dimensions two or three, but even then we can gain essential insight how the things may work in general (and this fact might be viewed as one of the strategies to build a mathematical mindset). Thus, we should enjoy the Sage tools designed to plot vectors and similar objects.<sup>1</sup>*

**2.A.5.** Use Sage for plotting: (a) The vectors  $u = (1, 2, 3)^T$ ,  $v = (2, -3, 4)^T$ ,  $u + v$ ,  $u - v$ , which are all vectors of the 3-dimensional Euclidean space  $\mathbb{R}^3$ .

(b) Repeat for the vectors  $x = (4, 0, 2)^T$ ,  $y = (-3, 0, 1)^T$ ,  $x + y$ , and  $x - y$ .

**Solution.** (a) The simplest method is to introduce the vectors, and then use the function `plot`, as follows:

```
u = vector([1, 2, 3])
```

<sup>1</sup>Notice that so far Sage allows us to introduce vectors and make operations between them, without a reference to the choice of scalars. Later in 2.C.12 we will see how to treat vectors in Sage, when they are viewed as elements of some specific vector space.

The properties (V1)–(V4) of our vectors are easily checked for any specific ring of scalars  $\mathbb{K}$ , since we need just the corresponding properties of scalars as listed in 1.1.1, applied to individual components of the vectors. In this way we shall work with, for instance,  $\mathbb{R}^n$ ,  $\mathbb{Q}^n$ ,  $\mathbb{C}^n$ , but also with  $\mathbb{Z}^n$ ,  $(\mathbb{Z}_k)^n$ ,  $n = 1, 2, 3, \dots$

**2.1.2. Matrices over scalars.** Matrices are slightly more complicated objects, useful when working with vectors.

#### MATRICES OF TYPE $m/n$

A matrix of the type  $m/n$  over scalars  $\mathbb{K}$  is a rectangular schema  $A$  with  $m$  rows and  $n$  columns

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

where  $a_{ij} \in \mathbb{K}$  for all  $1 \leq i \leq m$ ,  $1 \leq j \leq n$ . For a matrix  $A$  with elements  $a_{ij}$  we also use the notation  $A = (a_{ij})$ .

The vector  $(a_{i1}, a_{i2}, \dots, a_{in}) \in \mathbb{K}^n$  is called the ( $i$ -th) row of the matrix  $A$ ,  $i = 1, \dots, m$ . The vector  $(a_{1j}, a_{2j}, \dots, a_{mj}) \in \mathbb{K}^m$  is called the ( $j$ -th) column of the matrix  $A$ ,  $j = 1, \dots, n$ .

Matrices of the type  $1/n$  or  $n/1$  are actually just vectors in  $\mathbb{K}^n$ .

All general matrices can be understood as vectors in  $\mathbb{K}^{mn}$ , we just consider all the columns. In particular, matrix addition and matrix multiplication by scalars is defined:

$$A + B = (a_{ij} + b_{ij}), \quad a \cdot A = (a \cdot a_{ij})$$

where  $A = (a_{ij})$ ,  $B = (b_{ij})$ ,  $a \in \mathbb{K}$ .

The matrix  $-A = (-a_{ij})$  is called the additive inverse to the matrix  $A$  and the matrix

$$0 = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}$$

is called the zero matrix. By considering matrices as  $mn$ -dimensional vectors, we obtain the following:

**Proposition.** The formulas for  $A+B$ ,  $a \cdot A$ ,  $-A$ ,  $0$  define the operations of addition and multiplication by scalars for the set of all matrices of the type  $m/n$ , which satisfy properties (V1)–(V4).

**2.1.3. Matrices and equations.** Many mathematical models are based on systems of linear equations. Matrices are useful for the description of such systems. In order to see this, let us introduce the notion of scalar product of two vectors, assigning to the vectors  $(a_1, \dots, a_n)$  and  $(x_1, \dots, x_n)$  their product

$$(a_1, \dots, a_n) \cdot (x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n.$$

This means, we multiply the corresponding coordinates of the vectors and sum the results.

```
v = vector([2, -3, 4])
u_plus_v = u + v
u_minus_v = u - v
uu=plot(u, color="blue")
vv=plot(v, color="green")
uplusv=plot(u_plus_v, color="green")
uminusv=plot(u_minus_v, color="purple")
show(uu+vv+uplusv+uminusv)
```

However, Sage offers an alternative method for plotting vectors in three dimensions, using the `arrow3d` function. For instance, to visualize the vectors in (a) you can use the following block:

```
u=vector([1, 2, 3])
v=vector([2, -3, 4])
u_plus_v=u + v
u_minus_v=u - v
u_ar=arrow3d([0, 0, 0], u, color="red")
v_ar=arrow3d([0, 0, 0], v, color="blue")
u_plus_v_ar=arrow3d([0, 0, 0],
u_plus_v, color="green")
u_minus_v_ar=arrow3d([0, 0, 0],
u_minus_v, color="purple")
p=u_ar+v_ar+u_plus_v_ar+u_minus_v_ar
p.show()
```

In this block we used red, blue, green and purple to represent the vectors  $u, v, u+v$  and  $u-v$ , respectively. Run the code in your Sage environment to generate and view the vector plot. Next apply the same method to analyze the second set of vectors. □

*As our first real task essentially involving matrices, we are going to develop a systematic approach for solving systems of linear equations. Our first aim is to introduce the beautiful Gauss elimination method (see 2.1.7 for details), treating a few examples. You might also look into 2.1.3. Later, in Section C, we shall relate the description given here with the important notions of “affine spaces” and “vector spaces”.*

**2.A.6. A colourful example.** A company of painters orders 810 litres of paint, to contain 270 litres each of red, green and blue coloured paint. The provider can satisfy this order by mixing the colours he usually sells, namely:

- reddish colour – it contains 50% of red, 25% of green and 25% of blue colour;
- greenish colour – it contains 12.5% of red, 75% of green and 12.5% of blue colour;
- bluish colour – it contains 20% of red, 20% of green and 60% of blue colour.

How many litres of each of the colours at the warehouse have to be mixed in order to satisfy the order?

**Solution.** Let us denote by

- $x$  – the number of litres of reddish colour to be used;
- $y$  – the number of litres of bluish colour to be used;

Every system of  $m$  linear equations in  $n$  variables

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

can be seen as a constraint on values of  $m$  scalar products with one unknown vector  $(x_1, \dots, x_n)$  (called the *vector of variables*, or *vector variable*) and the known vectors of coordinates  $(a_{i1}, \dots, a_{in})$ .

The vector of variables can be also seen as a column in a matrix of the type  $n/1$ , and similarly the values  $b_1, \dots, b_m$  can be seen as a vector  $u$ , and that is again a single column of the matrix of the type  $m/1$ . Our system of equations can then be formally written as  $A \cdot x = u$  as follows:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

where the left-hand side is interpreted as  $m$  scalar products of the individual rows of the matrix (giving rise to a column vector) with the vector variable  $x$ , whose values are prescribed by the equations. That means that the identity of the  $i$ -th coordinates corresponds to the original  $i$ -th equation

$$a_{i1}x_1 + \cdots + a_{in}x_n = b_i$$

and the notation  $A \cdot x = u$  gives the original system of equations.

**2.1.4. Matrix product.** In the plane, that is, for vectors of dimension two, we developed a matrix calculus. We noticed that it is effective to work with (see 1.5.4). Now we generalize such a calculus and we develop all the tools we know already from the plane case to deal with higher dimensions  $n$ .

It is possible to define matrix multiplication only when the dimensions of the rows and columns allow it, that is, when the scalar product is defined for them as before:

MATRIX PRODUCT

For any matrix  $A = (a_{ij})$  of the type  $m/n$  and any matrix  $B = (b_{jk})$  of the type  $n/q$  over the ring of scalars  $\mathbb{K}$  we define their product  $C = A \cdot B = (c_{ik})$  as a matrix of the type  $m/q$  with the elements

$$c_{ik} = \sum_{j=1}^n a_{ij}b_{jk}, \text{ for arbitrary } 1 \leq i \leq m, 1 \leq k \leq q.$$

That is, the element  $c_{ik}$  of the product is exactly the scalar product of the  $i$ -th row of the matrix on the left and of the  $k$ -th column of the matrix on the right. For instance we have

$$\begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix}.$$

- $z$  – the number of litres of greenish colour to be used;

By mixing the colours, we want the final colour to contain 270 litres of red. By assumption, the reddish contains 50% of red, the greenish contains 12.5% of red and the bluish 20% of red. Hence, we get the equation

$$0.5x + 0.125y + 0.2z = 270.$$

Similarly, for blue and green colour, respectively, we get

$$\begin{aligned} 0.25x + 0.75y + 0.2z &= 270, \\ 0.25x + 0.125y + 0.6z &= 270. \end{aligned}$$

From the first equation we have  $x = 540 - 0.25y - 0.4z$ , and then the second and third equations give  $2.75y + 0.4z = 540$  and  $0.25y + 2z = 540$ , respectively. Hence,  $z = 270 - 0.125y$  and after substituting into the first equation, we obtain  $2.7y = 432$ , that is,  $y = 160$ . Therefore  $z = 270 - 0.125 \cdot 160 = 250$  and  $x = 540 - 0.25 \cdot 160 + 0.4 \cdot 250 = 400$ . This means that it is necessary to mix 400 litres of reddish, 160 litres of bluish and 250 litres of greenish colour.

Once you have set up the system that describes the problem, you can quickly find the solution using Sage. For example, enter the following block

```
var("x, y, z")
eq1=0.5*x+0.125*y+0.2*z-270
eq2=0.25*x+0.75*y+0.2*z-270
eq3=0.25*x+0.125*y+0.6*z-270
solve([eq1==0,eq2==0, eq3==0], x, y, z)
```

Sage's output is the list

```
[[x == 400, y == 160, z == 250]] □
```

**2.A.7. Matrix notation.** Approach the previous problem using the matrix notation and solving the matrix equation in Sage.



**Solution.** A practical approach to solve linear systems of equations relies on matrix notation. The first step is to construct the corresponding *coefficient matrix*, where each row represents the coefficients of the variables in a specific equation. In particular, the first row contains the coefficients from the first equation, the second row from the second equation, and so forth.

In our case, the system in 2.A.6 leads to the  $3 \times 3$  matrix

$$A = \begin{pmatrix} 0.5 & 0.125 & 0.2 \\ 0.25 & 0.75 & 0.2 \\ 0.25 & 0.125 & 0.6 \end{pmatrix}.$$

Then we write the system of equations as  $Ax = b$ , where  $x = (x_1, x_2, x_3)^T$  and  $b = (270, 270, 270)^T$ , respectively. We call  $x$  the vector of unknowns, and  $b$  the vector of constants in the right hand side of the system.

Merging  $A$  and the vector  $b$ , we arrive at the so called *extended matrix*<sup>2</sup>, which in our example has got the form

$$(A \mid b) = \left( \begin{array}{ccc|c} 0.5 & 0.125 & 0.2 & 270 \\ 0.25 & 0.75 & 0.2 & 270 \\ 0.25 & 0.125 & 0.6 & 270 \end{array} \right). \quad (\#)$$

<sup>2</sup>The extended matrix is also referred to as the *augmented matrix*.

**2.1.5. Square matrices.** If there is the same number of rows and columns in the matrix, we speak of a *square matrix*. The number of rows or columns is then called the *dimension of the matrix*. The matrix

$$E = (\delta_{ij}) = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix}$$

is called the *unit matrix*, or alternatively, the *identity matrix*. The numbers  $\delta_{ij}$  defined in such a way are also called the *Kronecker delta*. When we restrict ourselves to square matrices over  $\mathbb{K}$  of fixed dimension  $n$ , the matrix product is defined for any two matrices. That is, there is the well defined multiplication operation there. Its properties are similar to that of scalars:

**Proposition.** *On the set of all square matrices of dimension  $n$  over an arbitrary ring of scalars  $\mathbb{K}$ , the multiplication operation is defined with the following properties of rings (see 1.1.1):*

- (R1) Multiplication is associative.
- (R3) The unit matrix  $E = (\delta_{ij})$  is the unit element for multiplication.
- (R4) Multiplication and addition is distributive.

*In general, neither the property (R2) nor (ID) are true. Therefore, the square matrices for  $n > 1$  do not form an integral domain, and consequently they cannot be a (commutative or non-commutative) field.*

**PROOF.** Associativity of multiplication – (R1): Since scalars are associative, distributive and commutative, we can compute for any three matrices  $A = (a_{ij})$  of type  $m/n$ ,  $B = (b_{jk})$  of type  $n/p$  and  $C = (c_{kl})$  of type  $p/q$ :

$$\begin{aligned} A \cdot B &= \left( \sum_j a_{ij} \cdot b_{jk} \right), & B \cdot C &= \left( \sum_k b_{jk} \cdot c_{kl} \right), \\ (A \cdot B) \cdot C &= \left( \sum_k \left( \sum_j a_{ij} b_{jk} \right) c_{kl} \right) = \left( \sum_{j,k} a_{ij} b_{jk} c_{kl} \right), \\ A \cdot (B \cdot C) &= \left( \sum_j a_{ij} \left( \sum_k b_{jk} c_{kl} \right) \right) = \left( \sum_{j,k} a_{ij} b_{jk} c_{kl} \right). \end{aligned}$$

Note that while computing, we relied on the fact that it does not matter in which order are we performing the sums and products, that is, we were relying on the properties of scalars.

We can easily see that multiplication by a unit matrix has the property of a unit element:

$$A \cdot E = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \\ a_{m1} & \dots & a_{mm} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \dots & 1 \end{pmatrix} = A$$

and similarly from the left,

$$E \cdot A = A.$$

It remains to prove the distributivity of multiplication and addition. Again using the distributivity of scalars we can

Observe that the extended matrix completely encodes the original linear system of equations (just leaving out the variable names). We shall come back to the extended matrix in the next task.

In order to find a solution of  $Ax = b$  in Sage, we can use the command `A.solve_right(b)`. For our case:

```
A=matrix(QQ, [[0.5, 0.125, 0.2], [0.25,
0.75, 0.2], [0.25, 0.125, 0.6]])
b=vector([270, 270, 270])
x = A.solve_right(b); print(x)
```

Or, instead of `A.solve_right(b)`, we could type `A\b`, i.e.,

```
A=matrix(QQ, [[0.5, 0.125, 0.2],
[0.25, 0.75, 0.2], [0.25, 0.125, 0.6]])
b=vector([270, 270, 270])
A \ b #this is a Matlab like command
```

In the sequel we will refer to `A\b` as the “backslash” operator in Sage. If we are interested in a system of the form  $xA = b$ , in order to solve for  $x$  we use the command `A.solve_left(b)`. Let us notice that if the matrix equation does not have a solution, Sage returns an error (see also 2.A.20 for a similar case), while if it does have solutions, Sage returns just one. We shall come back to this phenomenon later.

**2.A.8. Using the echelon form.** Solve the previous problem again, this time using the transformation to the echelon form of the extended matrix.

**Solution.** The mathematical approach suggests to look for “equivalent” systems of equations with the same solution sets, perhaps much nicer than the original one. The crucial observation is that *elementary row transformations* allow us to transform the extended matrix into a *row echelon form*, see 2.1.7. Recall, the elementary row operations correspond to interchanging of any two rows, multiplying each element of a row by a non-zero number, or adding (a multiple of) a row to another one.

Dealing with an extended matrix  $B = (A \mid b)$ , in Sage the elementary row operations are implemented via the following functions:<sup>3</sup>

- |                                      |   |
|--------------------------------------|---|
| <code>B.swap.rows()</code>           | interchange two rows,                                 |
| <code>B.rescale.row()</code>         | scale a row by a factor,                              |
| <code>B.add_multiple_of_row()</code> | add a multiple of one row to another row and replace. |

Recall, a matrix is in row echelon form if it meets the following requirements:

- The first non-zero number from the left in a non-zero row, known as the *leading coefficient* or *pivot*, is always to the right of the leading coefficient of the row above it.
- Rows containing only zeros are located at the bottom of the matrix.

<sup>3</sup>In Sage there are similar functions encoding the column operations.

easily calculate for matrices  $A = (a_{ij})$  of the type  $m/n$ ,  $B = (b_{jk})$  of the type  $n/p$ ,  $C = (c_{jk})$  of the type  $n/p$ ,  $D = (d_{kl})$  of the type  $p/q$

$$\begin{aligned}
 A \cdot (B + C) &= \left( \sum_j a_{ij}(b_{jk} + c_{jk}) \right) \\
 &= \left( \left( \sum_j a_{ij}b_{jk} \right) + \left( \sum_j a_{ij}c_{jk} \right) \right) = A \cdot B + A \cdot C \\
 (B + C) \cdot D &= \left( \sum_k (b_{jk} + c_{jk})d_{kl} \right) \\
 &= \left( \left( \sum_k b_{jk}d_{kl} \right) + \left( \sum_k c_{jk}d_{kl} \right) \right) = B \cdot D + C \cdot D.
 \end{aligned}$$

As we have seen in 1.5.4, two matrices of dimension two do not necessarily commute: for example

$$\begin{aligned}
 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \\
 \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.
 \end{aligned}$$

This gives us immediately a counterexample to the validity of (R2) and (ID). For matrices of type  $1/1$  both axioms clearly hold, because the scalars itself have them. For matrices of greater dimension the counterexamples can be obtained similarly. Simply place the counterexamples for dimension 2 in their left upper corner, and select the rest to be zero. (Verify this on your own!) □

In the proof we have actually worked with matrices of more general types, thus we have proved the properties in greater generality:

ASSOCIATIVITY AND DISTRIBUTIVITY

Matrix multiplication is associative and distributive, that is,

$$\begin{aligned}
 A \cdot (B \cdot C) &= (A \cdot B) \cdot C \\
 A \cdot (B + C) &= A \cdot B + A \cdot C,
 \end{aligned}$$

whenever are all the given operations defined. The unit matrix is a unit element for multiplication (both from the right and from the left).

**2.1.6. Inverse matrices.** With scalars we can do the following: from the equation  $a \cdot x = b$  with a fixed invertible  $a$  we can express  $x = a^{-1} \cdot b$  for any  $b$ . We would like to be able to do this for matrices too. So we need to solve the problem – how to tell that such a matrix exists, and if so, how to compute it?



We say that  $B$  is the *inverse* of  $A$  if

$$A \cdot B = B \cdot A = E.$$

Then we write  $B = A^{-1}$ . From the definition it is clear that both matrices must be square and of the same dimension  $n$ . A matrix which has an inverse is called an *invertible matrix* or a *regular square matrix*.

These two conditions ensure that all entries in a column below a leading coefficient are zeros.

Paragraph 2.1.8 explains how to interpret the elementary row transformations as matrix multiplication. This presentation is often very convenient, and we shall come back to this later in Section E, see for example the task in 2.E.3.

We will now apply this method to solve the system posed in 2.A.6. Let  $R_1, R_2, R_3$  denote the rows of the matrix obtained at each step of the procedure. We begin by considering the extended matrix  $B = (A \mid b)$ , as outlined in relation (#) of 2.A.7, and apply the row operations  $R_1 \rightarrow 2R_1$ ,  $R_2 \rightarrow 4R_2$  and  $R_3 \rightarrow 4R_3$ . This yields

$$B = (A \mid b) \sim \left( \begin{array}{ccc|c} 1 & 0.25 & 0.4 & 540 \\ 1 & 3 & 0.8 & 1080 \\ 1 & 0.5 & 2.4 & 1080 \end{array} \right).$$

In Sage these operations are implemented by the block

```
B=matrix(QQ, [[0.5, 0.125, 0.2, 270], [0.25,
0.75, 0.2, 270], [0.25, 0.125, 0.6, 270]])
B.rescale_row(0,2) #multiply the 1st row
# by 2 (Sage labels rows as 0,1,...)
B.rescale_row(1,4) #the 2nd row by 4
B.rescale_row(2,4) #the 3rd row by 4
show(B)
```

where the last command is used to verify the matrix obtained along this first step. Notice, the first argument in these commands is the index of the row to be replaced, the second argument is the row to form a multiple of, and the final argument is the scale factor. Next, `B.add_multiple_of_row(k,l,c)` will replace then  $k$ th row  $R_k$  by  $R_k + cR_l$ , where  $R_l$  denotes the  $l$ th row and  $c$  is some (non-zero) constant.

Using the resulting matrix, we proceed with the row operations  $R_2 \rightarrow R_2 - R_1$  and  $R_3 \rightarrow R_3 - R_1$ , while keeping  $R_1$  untouched:

$$(A \mid b) \sim \left( \begin{array}{ccc|c} 1 & 0.25 & 0.4 & 540 \\ 0 & 2.75 & 0.4 & 540 \\ 0 & 0.25 & 2 & 540 \end{array} \right).$$

These row operations also admit an easy description in Sage, simply by adding in the previous cell the code

```
B.add_multiple_of_row(1,0,-1)
#replace the 2nd row R2 by R2-R1
B.add_multiple_of_row(2,0,-1)
#replace the 3rd row R3 by R3-R1
show(B)
```

Notice, in Sage we do *not* need to introduce new matrices when doing row operations successively. Next we multiply both  $R_2$  and  $R_3$  by 4 and then interchange them, i.e.,  $R_1 \rightarrow R_1$ ,  $R_2 \rightarrow 4R_2$ ,  $R_3 \rightarrow 4R_3$  and next  $R_2 \leftrightarrow R_3$ . This provides

$$B \sim \left( \begin{array}{ccc|c} 1 & 0.25 & 0.4 & 540 \\ 0 & 11 & 1.6 & 2160 \\ 0 & 1 & 8 & 2160 \end{array} \right) \sim \left( \begin{array}{ccc|c} 1 & 0.25 & 0.4 & 540 \\ 0 & 1 & 8 & 2160 \\ 0 & 11 & 1.6 & 2160 \end{array} \right).$$

In order to describe this step in Sage, we should continue typing in the previous cell the code given below:

In the subsequent paragraphs we derive (among other things) that  $B$  is actually the inverse of  $A$  whenever just one of the above required equations holds. The other is then a consequence.

We easily check that if  $A^{-1}$  and  $B^{-1}$  exist, then there also is the inverse of the product  $A \cdot B$

$$(1) \quad (A \cdot B)^{-1} = B^{-1} \cdot A^{-1}.$$

Indeed, because of the associativity of matrix multiplication proved a while ago, we have

$$(B^{-1} \cdot A^{-1}) \cdot (A \cdot B) = B^{-1} \cdot (A^{-1} \cdot A) \cdot B = E$$

$$(A \cdot B) \cdot (B^{-1} \cdot A^{-1}) = A \cdot (B \cdot B^{-1}) \cdot A^{-1} = E.$$

Because we can calculate with matrices similarly as with scalars (they are just a little more complicated), the existence of an inverse matrix can really help us with the solution of systems of linear equations: if we express a system of  $n$  equations for  $n$  unknowns as a matrix product



$$A \cdot x = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = u$$

and when the inverse of the matrix  $A$  exists, then we can multiply from the left by  $A^{-1}$  to obtain

$$A^{-1} \cdot u = A^{-1} \cdot A \cdot x = E \cdot x = x,$$

that is,  $A^{-1} \cdot u$  is the desired solution.

On the other hand, expanding the condition  $A \cdot A^{-1} = E$  for unknown scalars in the matrix  $A^{-1}$  gives us  $n$  systems of linear equations for the same matrix on the left and different vectors on the right. Thus we should think about methods for solutions of the systems of linear equations.

**2.1.7. Equivalent operations with matrices.** Let us gain some practical insight into the relation between systems of equations and their matrices. Clearly, searching for the inverse can be more complicated than finding the direct solution to the system of equations. But note that whenever we have to solve more systems of equations with the same matrix  $A$  but with different right sides  $u$ , then yielding  $A^{-1}$  can be really beneficial for us.

From the point of view of solving systems of equations  $A \cdot x = u$ , it is natural to consider the matrices  $A$  and vectors  $u$  equivalent whenever they give a system of equations with the same solution set. Let us think about possible operations which would simplify the matrix  $A$  such that obtaining the solution is easier.



We begin with simple manipulations of rows of equations which do not influence the solution, and similar modifications of the right-hand side vector. If we are able to change a square matrix into the unit matrix, then the right-hand side vector is a solution of the original system. If some of the rows of the system vanish during the course of manipulations (that is,

```
B.rescale_row(1,4)
#multiply the 2nd row R2 by 4
B.rescale_row(2,4)
#multiply the 3rd row R3 by 4
B.swap_rows(1,2)
#interchange the 2nd row by the 3rd row
show(B)
```

Observe now that in the matrix obtained in this step, the coefficient of  $y$  in the third row is not yet zero. Therefore, it remains to apply one more row operation, namely  $R_3 \rightarrow R_3 - 11R_2$ :

$$B \sim \left( \begin{array}{ccc|c} 1 & 0.25 & 0.4 & 540 \\ 0 & 1 & 8 & 2160 \\ 0 & 0 & -86.4 & -21600 \end{array} \right).$$

This matrix is in row echelon form, and in Sage for this last row operation we successively proceed with the code

```
B.add_multiple_of_row(2,1,-11)
#replace the 3rd row R3 by R3-11R2
show(B)
```

Finally, by the so called *backward substitution*, we can describe the solution of the initial system. In particular, from the last row we get  $z = \frac{-21600}{-86.4} = 250$ . Using this, from the second row we obtain  $y = 2160 - 8 \cdot 250 = 160$  and finally from the first row we have  $x = 540 - 0.4 \cdot 250 - 0.25 \cdot 160 = 400$ .  $\square$

**2.A.9.** Based on elementary row transformation solve the linear system posed below. Then use Sage to find an echelon form of the corresponding extended matrix, and verify your solution:

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 2, \\ 2x_1 - 3x_2 - x_3 = -3, \\ -3x_1 + x_2 + 2x_3 = -3. \end{cases}$$

**Solution.** Let us express the system in the matrix form  $Ax = b$ :

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & -3 & -1 \\ -3 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \\ -3 \end{pmatrix},$$

. The extended matrix is given by

$$B := (A | b) = \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 2 & -3 & -1 & -3 \\ -3 & 1 & 2 & -3 \end{array} \right).$$

Let us now perform elementary row operations to transform the extended matrix into row echelon form. We have:

$$\begin{aligned} (A | b) &\sim \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -7 & -7 & -7 \\ 0 & 7 & 11 & 3 \end{array} \right) \\ &\sim \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & -7 & -7 & -7 \\ 0 & 0 & 4 & -4 \end{array} \right) \sim \left( \begin{array}{ccc|c} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{array} \right). \end{aligned}$$

Above, we first applied the operations  $R_2 \rightarrow R_2 - 2R_1$ ,  $R_3 \rightarrow R_3 + 3R_1$  on  $B$  to obtain the first mentioned matrix. Using this, we proceeded with  $R_3 \rightarrow R_3 + R_2$ , to get the

they become zero), then we get some direct information about the solution. Our simple operations are:

#### ELEMENTARY ROW TRANSFORMATIONS

- interchanging two rows,
- multiplication of any given row by a non-zero scalar,
- adding another row to any given row.

These operations are called *elementary row transformations*. It is clear that the corresponding operations at the level of the equations in the system do not change the set of the solutions whenever we deal with a field of scalars, they still look reasonable if our ring of coordinates is an integral domain (but multiplying by a divisor of zero might be a problem).

Analogically, *elementary column transformations* of matrices are

- interchanging two columns
- multiplication of any given column by a non-zero scalar,
- adding another column to any given column.

These do not preserve the solution set, since they change the variables themselves.

Systematically we can use elementary row transformations for subsequent elimination of variables. This gives an algorithm which is usually called the *Gaussian elimination method*. Henceforth, we shall assume that our scalars come from an integral domain (e.g. integers are allowed, but not say  $\mathbb{Z}_4$ ).



#### GAUSSIAN ELIMINATION OF VARIABLES

**Proposition.** Any non-zero matrix over an arbitrary integral domain of scalars  $\mathbb{K}$  can be transformed, using finitely many elementary row transformations, into row echelon form:

- For each  $j$ , if  $a_{ik} = 0$  for all columns  $k = 1, \dots, j$ , then  $a_{kj} = 0$  for all  $k \geq i$ ,
- if  $a_{(i-1)j}$  is the first non-zero element at the  $(i-1)$ -st row, then  $a_{ij} = 0$ .

**PROOF.** The matrix in row echelon form looks like

$$\begin{pmatrix} 0 & \dots & 0 & a_{1j} & \dots & \dots & \dots & a_{1m} \\ 0 & \dots & 0 & 0 & \dots & a_{2k} & \dots & a_{2m} \\ \vdots & & & & & & & \\ 0 & \dots & \dots & \dots & \dots & 0 & a_{lp} & \dots \\ \vdots & & & & & & & \end{pmatrix}.$$

The matrix can (but does not have to) end with some zero rows. In order to transform an arbitrary matrix, we can use a simple algorithm, which will bring us, row by row, to the resulting echelon form:

second matrix, and the final step consists of the operations  $R_2 \rightarrow -\frac{1}{7}R_2$  and  $R_3 \rightarrow \frac{1}{4}R_3$ , which give the third matrix. This one is in row echelon form and implies directly  $x_3 = -1$ , and moreover

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 2, \\ x_2 + x_3 = 1, \end{cases}$$

from where we specify  $x_2 = 2$  and  $x_1 = 1$ .

Sage provides an easy way to compute a row echelon form corresponding to a matrix  $B$ , via the command `B.echelon_form()`. For our example we can type

```
A=matrix([[1, 2, 3], [2, -3, -1], [-3, 1, 2]])
b=vector([2, -3, -3])
B=A.augment(b) #the augmented matrix
show(B)
show(B.echelon_form())
```

This code returns the extended matrix  $B$  and a row echelon form associated to  $B$ :<sup>4</sup>

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & 7 & 3 & 11 \\ 0 & 0 & 4 & -4 \end{pmatrix}.$$

Note that Sage’s solution differs from ours. However, there is no need to worry, as there are various methods for performing row reduction, and thus the row echelon form of a matrix is *not* unique! From the last row of this matrix we obtain  $4x_3 = -4$ , that is,  $x_3 = -1$ , and so the second row which reads as  $7x_2 + 3x_3 = 11$ , becomes  $7x_2 = 14$ , so  $x_2 = 2$ . From the first row we have  $x_1 + 2x_2 + 3x_3 = 2$ , and by replacing the values found for  $x_2, x_3$  we get  $x_1 = 1$ , and so the same solution as above.  $\square$

*In summary, the most efficient method for solving large systems of linear equations involves applying elementary row transformations to achieve row echelon form. This technique, known as Gaussian elimination, can be optimized through various sophisticated schemes for selecting the best pivots in numerical contexts. Subsequent examples will illustrate that systems of linear equations may have infinitely many solutions, exactly one solution, or none at all. For a theoretical exploration of the classification of solutions using matrix invariants and other methods, see 2.1.13 and 2.3.5.*

**2.A.10.** Using the Gauss elimination method solve the following linear system:

$$\begin{cases} 2x_1 - x_2 + 3x_3 = 0, \\ 3x_1 + 16x_2 + 7x_3 = 0, \\ 3x_1 - 5x_2 + 4x_3 = 0, \\ -7x_1 + 7x_2 + -10x_3 = 0. \end{cases}$$

<sup>4</sup>If we want the vertical line in the presentation of the extended matrix, we can type `B = A.augment(b, subdivide = "True")`.

GAUSSIAN ELIMINATION ALGORITHM

- (1) By a possible interchange of rows we can obtain a matrix where the first row has a non-zero element in the first non-zero column. Let that column be column  $j$ . In other words,  $a_{1j} \neq 0$ , but  $a_{iq} = 0$  for all  $i$ , and all  $q$ ,  $1 \leq q < j$ .
- (2) For each  $i = 2, \dots$ , multiply the first row by the element  $a_{ij}$ , multiply  $i$ -th row by the element  $a_{1j}$  and subtract, to obtain  $a_{ij} = 0$  on the  $i$ -th row.
- (3) By repeated application of the steps (1) and (2), always for the not-yet-echelon part of rows and columns in the matrix we reach, after a finite number of steps, the final form of the matrix.

This algorithm clearly stops after a finite number of steps and provides the proof of the proposition.  $\square$

The given algorithm is really the usual elimination of variables used in the systems of linear equations.

In a completely analogous manner we define the column echelon form of matrices and considering column elementary transformations instead the row ones, we obtain an algorithm for transforming matrices into the column echelon form.

**Remark.** Although we could formulate the Gaussian elimination for general scalars from any ring, this does not make much sense in view of solving equations. Clearly having divisors of zero among the scalars, we might get zeros during the procedure and lose information this way. Think carefully about the differences between the choices  $\mathbb{K} = \mathbb{Z}$ ,  $\mathbb{K} = \mathbb{R}$  and possibly  $\mathbb{Z}_2$  or  $\mathbb{Z}_4$ .



On the other hand, if we are dealing with fields of scalars, we can always arrive at a row echelon form where the non-zero entries on the “diagonal” are ones. This is done by applying the appropriate scalar multiplication to each individual row. However, this is not possible in general – think for instance of the integers  $\mathbb{Z}$ .

**2.1.8. Matrices of elementary row transformations.** Let us now restrict ourselves to fields of scalars  $\mathbb{K}$ , that is, every non-zero scalar has an inverse.

Note that elementary row or column transformations correspond respectively to multiplication from the left or right by the following matrices (only the differences from the unit matrix are indicated):

- (1) Interchanging the  $i$ -th and  $j$ -th row (column)

$$\begin{pmatrix} \ddots & & & & \\ & 0 & \dots & 1 & \\ & \vdots & \ddots & \vdots & \\ & 1 & \dots & 0 & \\ & & & & \ddots \end{pmatrix} \begin{array}{l} \leftarrow i\text{-th row} \\ \leftarrow j\text{-th row} \end{array}$$



**Solution.** This is a so called *homogeneous linear system*, since the vector  $b$  is the zero vector. In matrix notation it reads as

$$\begin{pmatrix} 2 & -1 & 3 \\ 3 & 16 & 7 \\ 3 & -5 & 4 \\ -7 & 7 & -10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

that is,  $Ax = 0$ , where  $0$  denotes the zero vector. A homogeneous system of linear equations has either exactly one solution, the zero one, or infinitely many solutions, see 2.3.5 for details. By transforming the matrix into a row echelon form via elementary row operations, we will prove that the given system has infinitely many solutions. Initially we obtain

$$A \sim \begin{pmatrix} 2 & -1 & 3 \\ 3 & 16 & 7 \\ 3 & -5 & 4 \\ -7 & 7 & -10 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & -7/2 & -1/2 \\ 0 & 7/2 & 1/2 \end{pmatrix}.$$

From there, we see that the second, third and fourth equations are multiples of the equation  $7x_2 + x_3 = 0$ . We continue as follows:

$$\begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & -7/2 & -1/2 \\ 0 & 7/2 & 1/2 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 35/2 & 5/2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 2 & -1 & 3 \\ 0 & 7 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

We can translate this literally as

$$\left\{ 2x_1 - x_2 + 3x_3 = 0, \quad 7x_2 + x_3 = 0 \right\},$$

where observe that the equations induced by the last two rows are redundant. In particular,  $x_3$  is a free variable to which we assign the real parameter  $t \in \mathbb{R}$ . This gives infinite many solutions of the form

$$x_2 = -\frac{1}{7}x_3 = -\frac{1}{7}t, \quad x_1 = \frac{1}{2}(x_2 - 3x_3) = -\frac{11}{7}t.$$

We may verify this result in Sage, as before:

```
var("x1, x2, x3"); eq1=2*x1-x2+3*x3
eq2=3*x1+16*x2+7*x3; eq3=3*x1-5*x2+4*x3
eq4=-7*x1+7*x2-10*x3
solve([eq1==0, eq2==0, eq3==0, eq4==0],
x1, x2, x3)
```

which answers

```
x1 == -11/7*r1, x2 == -1/7*r1, x3 == r1
```

Notice the substitution  $t = -7s$  brings our solution to the form  $(11s, s, -7s)$ , with  $s \in \mathbb{R}$ , that is,

$$(x_1, x_2, x_3)^T = s \cdot (11, 1, -7)^T, \quad s \in \mathbb{R},$$

where “ $\cdot$ ” is simply the scalar multiplication. This expression can be obtained very quickly in Sage, by the command `A.right_kernel()`, where  $A$  is the matrix of the linear homogeneous system, as discussed in the next task.  $\square$

(2) Multiplication of the  $i$ -th row (column) by the scalar  $a$ :

$$\begin{pmatrix} \ddots & & & & \\ & 1 & & & \\ & & a & & \\ & & & 1 & \\ & & & & \ddots \end{pmatrix} \leftarrow i\text{-th row}$$

(3) To row  $i$ , add row  $j$  (columns):

$$i\text{-th row and } j\text{-th column} \rightarrow \begin{pmatrix} \ddots & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & \ddots \end{pmatrix}$$

This trivial observation is actually very important, since the product of invertible matrices is invertible (recall 2.1.6(1)) and all elementary transformations over a field of scalars are invertible (the definition of the elementary transformation itself ensures that inverse transformations are of the same type and it is easy to determine the corresponding matrix).

Thus, the Gaussian elimination algorithm tells us, that for an arbitrary matrix  $A$ , we can obtain its equivalent row echelon form  $A' = P \cdot A$  by multiplying with a suitable invertible matrix  $P = P_k \cdots P_1$  from the left (that is, sequential multiplication with  $k$  matrices of the elementary row transformations).

If we apply the same elimination procedure for the columns, we can transform any matrix  $B$  into its column echelon form  $B'$  by multiplying it from the right by a suitable invertible matrix  $Q = Q_1 \cdots Q_\ell$ . If we start with the matrix  $B = A'$  in row echelon form, this procedure eliminates only the still non-zero elements out of the “diagonal” of the matrix and in the end we can transform the remaining elements to be units. Thus we have verified a very important result which we will use many times in the future:

**2.1.9. Theorem.** *For every matrix  $A$  of the type  $m/n$  over a field of scalars  $\mathbb{K}$ , there exist square invertible matrices  $P$  and  $Q$  of dimensions  $m$  and  $n$ , respectively, such that the matrix  $P \cdot A$  is in row echelon form and*

$$P \cdot A \cdot Q = \begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & & & & \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & & \end{pmatrix}.$$

*The number of the ones in the diagonal is independent of the particular choice of  $P$  and  $Q$ .*

**PROOF.** We already have proved everything but the last sentence. We shall see this last claim below in 2.1.11.  $\square$

**2.A.11.** Analyze the space of solutions in the previous task in view of the vector addition and scalar multiplication.

**Solution.** Of course, knowing that  $Ax = 0$ ,  $Ay = 0$ , the sum must be a solution again:  $A(x + y) = Ax + Ay = 0$ . Similarly  $A(cx) = c(Ax)$  for all scalars  $c$ . Viewing matrices as linear mappings  $x \mapsto Ax$ , the *kernel* of a (real)  $m \times n$  matrix  $A$ , also known as the *null space* of  $A$ , is the set of vectors  $x \in \mathbb{R}^n$  satisfying  $Ax = 0$ , where  $0$  denotes the zero vector of degree  $m$ .

We saw in the previous task that the solution space, i.e., the kernel of  $A$ , was spanned by one vector  $(11, 1, -7)^T$ .

To compute the null space of  $A$  in Sage we can just type `A.right_kernel()`. However, be aware that Sage also provides the command `A.kernel()`, which refers to the left kernel of  $A$  and it is an alias of the command `A.left_kernel()`. In our text we will use only the left/right kernel commands to avoid confusions.

Now, to obtain a verification of the previous expression you can type

```
A=matrix([[2, -1, 3], [3, 16, 7],
          [3, -5, 4], [-7, 7, -10]])
sol=A.right_kernel()
show(sol)
```

The output has the form `RowSpan(11 1 -7)`, which essentially means that the corresponding solution is spanned by the vector  $(11, 1, -7)^T$ , as we described above. Alternatively, you can simply type `A.right_kernel()`, and then Sage returns a more complex information:

```
Free module of degree 3 and rank 1
over Integer Ring
Echelon basis matrix: [11 1 -7]
```

In the first line of this output, the “degree 3” part refers to the “dimension” of the vectors in the kernel, that is  $3 \times 1$  vectors. The “rank 1” part indicates the “dimension” of the right kernel of  $A$ , i.e., of the null space of  $A$ .

Notice, Sage automatically considered  $A$  over the ring  $\mathbb{Z}$ , and therefore it talks about *free modules* instead of *vector spaces*. The answer would be different if we enforced the field  $\mathbb{Q}$  (by adding `QQ` in the definition of  $A$ ).<sup>5</sup>

We will learn about vector spaces, subspaces and the notion of dimension of a vector space in Section C, where we will meet further examples on null spaces. □

**2.A.12.** Based on Gauss elimination method, solve the linear system below. Next present the solution in Sage.

$$\begin{cases} 3x_1 & & + & 3x_3 & - & 5x_4 & = & -8, \\ x_1 & - & x_2 & + & x_3 & - & x_4 & = & -2, \\ -2x_1 & - & x_2 & + & 4x_3 & - & 2x_4 & = & 0, \\ 2x_1 & + & x_2 & - & x_3 & - & x_4 & = & -3. \end{cases}$$

<sup>5</sup>Beware that declaring  $A$  to be over the field  $\mathbb{R}$ , we obtain a completely different result since the numerical errors in the 53 bit arithmetic prevent us to have the echelon form with two zero rows. Thus Sage would answer that there is only the zero solution there!

**2.1.10. Algorithm for computing inverse matrices.** In the previous paragraphs we almost obtained the complete algorithm for computing the inverse matrix. Using the simple modification below, we find either that the inverse does not exist, or we compute the inverse. Keep in mind that we are still working over a field of scalars.



Equivalent row transformations of a square matrix  $A$  of dimension  $n$  leads to an invertible matrix  $P'$  such that  $P' \cdot A$  is in row echelon form. If  $A$  has an inverse, then there exists also the inverse of  $P' \cdot A$ . But if the last row of  $P' \cdot A$  is zero, then the last row of  $P' \cdot A \cdot B$  is also zero for any matrix  $B$  of dimension  $n$ . Thus, the existence of a zero row in the result of (row) Gaussian elimination excludes the existence of  $A^{-1}$ .

Assume now that  $A^{-1}$  exists. As we have just seen, the row echelon form of  $A$  will have exclusively non-zero rows only. In particular, all diagonal elements of  $P' \cdot A$  are non-zero. But now, we can employ row elimination by the elementary row transformation from the bottom-right corner backwards and also transform the diagonal elements to be units. In this way, we obtain the unit matrix  $E$ . Summarizing, we find another invertible matrix  $P''$  such that for  $P = P'' \cdot P'$  we have  $P \cdot A = E$ .

Now observe that we could clearly work with columns instead of row transformation and thus, under the assumption of the existence of  $A^{-1}$ , we would find a matrix  $Q$  such that  $A \cdot Q = E$ . From this we see immediately that

$$P = P \cdot E = P \cdot (A \cdot Q) = (P \cdot A) \cdot Q = E \cdot Q = Q.$$

That is, we have found the inverse matrix

$$A^{-1} = P = Q$$

for the matrix  $A$ . Notice that at the point of finding the matrix  $P$  with the property  $P \cdot A = E$ , we do not have to do any further computation, since we have already obtained the inverse matrix.

In practice, we can work as follows:

COMPUTING THE INVERSE MATRIX

Write the unit matrix  $E$  to the right of the matrix  $A$ , producing an augmented matrix  $(A, E)$ . Transform the augmented matrix using the elementary row transformations to row echelon form. This produces an augmented matrix  $(PA, PE)$ , where  $P$  is invertible, and  $PA$  is in row echelon form. By the above, either we may achieve  $PA = E$ , in which case  $A$  is invertible and  $P = PE = A^{-1}$ , or  $PA$  has a row of zeros, in which case we conclude that the inverse matrix for  $A$  does not exist.

**2.1.11. Linear dependence and rank.** In the previous practical algorithms dealing with matrices we worked all the time with row and column additions and scalar multiplications, seeing them as vectors.



Such operations are called *linear combinations*. We shall return to such operations in an abstract sense later on in 2.3.1.

**Solution.** The corresponding extended matrix has the form

$$B := (A \mid b) = \left( \begin{array}{cccc|c} 3 & 0 & 3 & -5 & -8 \\ 1 & -1 & 1 & -1 & -2 \\ -2 & -1 & 4 & -2 & 0 \\ 2 & 1 & -1 & -1 & -3 \end{array} \right).$$

By changing the order of rows (equations) we obtain

$$B \sim \left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 2 & 1 & -1 & -1 & -3 \\ -2 & -1 & 4 & -2 & 0 \\ 3 & 0 & 3 & -5 & -8 \end{array} \right)$$

which we transform into a row echelon form:

$$\begin{aligned} \left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 2 & 1 & -1 & -1 & -3 \\ -2 & -1 & 4 & -2 & 0 \\ 3 & 0 & 3 & -5 & -8 \end{array} \right) &\sim \left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & -3 & 6 & -4 & -4 \\ 0 & 3 & 0 & -2 & -2 \end{array} \right) \\ &\sim \left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 3 & -3 & -3 \end{array} \right) &\sim \left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right). \end{aligned}$$

Notice that via the command `B.echelon_form()`, Sage computes a different echelon form. In any case, we obtain three equations in four variables, and thus the initial linear system has infinitely many solutions (as in the previous example, although this is not a homogeneous system). These three equations have exactly one solution for any choice for the free variable  $x_4 \in \mathbb{R}$ . So we may set  $x_4 = t \in \mathbb{R}$ , and translate the final matrix of the Gauss elimination as

$$\begin{aligned} x_1 - x_2 + x_3 - t &= -2, \\ 3x_2 - 3x_3 + t &= 1, \\ 3x_3 - 3t &= -3. \end{aligned}$$

The solution now can be easily described by back substitution:

$$x_3 = t - 1, \quad x_2 = \frac{1}{3}(2t - 2), \quad x_1 = \frac{1}{3}(2t - 5).$$

Of course, the same solution occurs if we use the echelon form that Sage provides. Now, to have a Sage verification of the given solution, it sufficient to use the cell

```
Eq1=x1-x2+x3-x4+2
Eq2=3*x2-3*x3+x4-1
Eq3=3*x3-3*x4+3
show(solve([Eq1==0, Eq2==0, Eq3==0],
           x1, x2, x3))
```

Notice for  $t = 3s$  our solution  $(x_1, x_2, x_3, x_4)$  takes the form

$$\left( 2s - \frac{5}{3}, 2s - \frac{2}{3}, 3s - 1, 3s \right), \quad s \in \mathbb{R}. \quad \square$$

**2.A.13. Sage and general systems.** Analyze the solutions of the system of linear equations from the previous task in view of the vector operations and Sage tools.

**Solution.** We could approach the linear system of 2.A.12 by using the backslash operator in Sage:

But it will be useful to understand their core meaning right now. A linear combination of rows of a matrix  $A = (a_{ij})$  of type  $m/n$  is understood as an expression of the form

$$c_1 u_{i_1} + \dots + c_k u_{i_k},$$

where  $c_i$  are scalars,  $u_j = (a_{j1}, \dots, a_{jn})$  are rows of the matrix  $A$ . Similarly, we can consider linear combinations of columns by replacing the above rows  $u_j$  by the columns  $u_j = (a_{1j}, \dots, a_{mj})$ .

If the zero row can be written as a linear combination of some given rows with at least one non-zero scalar coefficient, we say that these rows are *linearly dependent*. In the alternative case, that is, when the only possibility of obtaining the zero row is to select all the scalars  $c_j$  equal to zero, the rows are called *linearly independent*.

Analogously, we define linearly dependent and linearly independent columns.

**Lemma.** *The number of non-zero “steps” in the row echelon form always equals to the number of linearly independent rows of the matrix. In particular, the elementary row transformations do not change the maximal number of linearly independent rows. Similarly for columns and the column echelon form, and the maximal number of linearly independent rows and columns coincides (and equals to the number of ones in Theorem 2.1.9)*



**PROOF.** Indeed, the claims all follow once we prove that the number  $h$  of ones in the matrix  $E_h$  from the theorem 2.1.9 does not depend on our choice of the row or column transformations in our procedure. Thus, assume that by two different row transformation procedures into the echelon form we obtain two different  $h' < h$ . But then according to our algorithm there are invertible matrices  $P', P'', Q',$  and  $Q''$  such that

$$E_h = P' \cdot A \cdot Q', \quad E_{h'} = P'' \cdot A \cdot Q''.$$

In particular,  $E_h = P' \cdot P''^{-1} \cdot E_{h'} \cdot Q''^{-1} \cdot Q'$  and so there are invertible matrices  $P$  and  $Q$  such that

$$E_{h'} \cdot Q = P \cdot E_h.$$

Now, on the left hand side, there are the first  $h'$  rows of  $Q$ , complemented by zero rows. At the same time, on the right hand side, there are  $h$  first columns of  $P$ , complemented by zero columns. Thus, the equality implies that taking the first  $h$  columns in  $P$ , only the first  $h'$  of the rows there are non-zero. But this shows, that our Gaussian elimination algorithm applied to  $P$  would show  $P$  is not invertible. Therefore  $h' = h$ , and we have proved that the number of ones in the matrix  $P \cdot A \cdot Q$  in theorem 2.1.9 is independent of the choice of our elimination procedure and it is always equal to the number of linearly independent rows in  $A$ , which must be the same as the number of linearly independent columns in  $A$ .  $\square$

**Definition.** The maximal number  $h(A)$  of linearly independent rows (or columns) in  $A$  is called the *rank of the matrix*.



We have the following theorem:

```
A=matrix([[3, 0, 3, -5], [1, -1, 1, -1],
          [-2, -1, 4, -2], [2, 1, -1, -1]])
b=vector([-8, -2, 0, -3])
x=A.solve_right(b); x
```

The output is the solution given above with  $t = 0$  only:

```
(-5/3, -2/3, -1, 0)
```

As we have seen before, this provides just one solution of the system. But the matrix form  $Ax = b$  of the equation reveals that having two such solutions  $x$  and  $y$ , ensures that  $A(x - y) = Ax - Ay = b - b = 0$  (notice we can compute with vectors and matrices just like if they were scalars). On the contrary, if  $y$  was a solution of the homogeneous system,  $Ay = 0$ , then  $A(x + y) = b + 0 = b$ .

Hence we should be grateful to Sage, picking one of the solutions to the original system and combining the backslash operator with the `right_kernel()` method, as follows:

```
A=matrix([[3, 0, 3, -5], [1, -1, 1, -1],
          [-2, -1, 4, -2], [2, 1, -1, -1]])
b=vector([-8, -2, 0, -3])
x=A.solve_right(b)
k=A.right_kernel().matrix()
# basis of solutions in rows
nrows = k.nrows()
# counts the kernel dimension
if nrows > 0: # non unique solution
    t=vector( var("t", n=nrows))
    # row of parameters t0, ...
    show(x + t*k)
else: # unique solution
    show(x)
```

Executing this block, Sage’s output is as in the end of the previous task. Actually, we add the vector subspace of the solutions to the homogeneous system, to one of the general solutions.  $\square$

*If we work over a field of scalars, e.g.,  $\mathbb{Q}$  or  $\mathbb{R}$ , the extended matrices  $B = (A \mid b)$  of the linear systems can be transformed further, aiming at having all the pivots equal to one, and all the other entries in the columns of the pivots equal to zero. This matrix is in a special row echelon form, known as the reduced row echelon form (often referred to as RREF).*

*It is not difficult to prove that this reduced row echelon form of any matrix is unique, once the pivot columns are fixed.*

**2.A.14.** Come back to the task in 2.A.12 and solve it using the RREF method. Then check with Sage.

**Solution.** A few elementary row transformations of the augmented matrix lead to

$$\left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 3 & -3 & 1 & 1 \\ 0 & 0 & 3 & -3 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right) \sim \left( \begin{array}{cccc|c} 1 & -1 & 1 & -1 & -2 \\ 0 & 1 & -1 & 1/3 & 1/3 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right)$$

**Theorem.** Let  $A$  be a matrix of type  $m/n$  over a field of scalars  $\mathbb{K}$ . The matrix  $A$  has the same maximal number  $h(A)$  of linearly independent rows as linearly independent columns. In particular, the rank is always at most the minimum of the dimensions of the matrix  $A$ .

The algorithm for computing the inverse matrix also says that a square matrix  $A$  of dimension  $m$  has an inverse if and only if its rank equals  $m$ .

**2.1.12. Matrices as mappings.** Similarly to the way we worked with matrices in the geometry of the plane (see 1.5.7), we can interpret every matrix  $A$  of the type  $m/n$  as a mapping

$$A : \mathbb{K}^n \rightarrow \mathbb{K}^m, \quad x \mapsto A \cdot x.$$

By the distributivity of matrix multiplication, it is clear how the linear combinations of vectors are mapped using such mappings:

$$A \cdot (ax + by) = a(A \cdot x) + b(A \cdot y).$$

Straight from the definition we see, by the associativity of multiplication, that composition of mappings corresponds to matrix multiplication in given order. Thus invertible matrices of dimension  $n$  correspond to bijective mappings

$$A : \mathbb{K}^n \rightarrow \mathbb{K}^n.$$

**Remark.** From this point of view, the theorem 2.1.9 is very interesting. We can see it as follows: the rank of the matrix determines how large is the image of the whole  $\mathbb{K}^n$  under this mapping. In fact, if  $A = P \cdot E_k \cdot Q$  where the matrix  $E_k$  has  $k$  ones as in 2.1.9, then the invertible  $Q$  first bijectively “shuffles” the  $n$ -dimensional vectors in  $\mathbb{K}^n$ , the matrix  $E_k$  then “copies” the first  $k$  coordinates and completes them with the remaining  $m - k$  zeros.

This “ $k$ -dimensional” image then cannot be enlarged by multiplying with  $P$ . Multiplying by  $P$  can only bijectively reshuffle the coordinates in the image.

**2.1.13. Back to linear equations.** We shall return to the notions of dimension, linear independence and so on in the third part of this chapter. But we should notice now what our results say about the solutions of the systems of linear equations.

If we consider the matrix of the system of equations and add to it the column of the required results, we speak about the *extended matrix of the system*. The above Gaussian elimination approach corresponds to the sequential variable elimination in the equations and the deletion of the linearly dependent equations (these are simply consequences of other equations).

Thus we have derived complete information about the size of the set of solutions of the system of linear equations, based on the rank of the matrix of the system. If we are left with more non-zero rows in the row echelon form of the extended matrix than in the original matrix of the system, then there cannot be a solution (simply, we cannot obtain the given vector value with the corresponding linear mapping). If the



$$\sim \left( \begin{array}{cccc|c} 1 & -1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -2/3 & -2/3 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right) \sim \left( \begin{array}{cccc|c} 1 & 0 & 0 & -2/3 & -5/3 \\ 0 & 1 & 0 & -2/3 & -2/3 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right).$$

The final matrix is in reduced row echelon form, which evidently directly provides the solution

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -5/3 \\ -2/3 \\ -1 \\ 0 \end{pmatrix} + t \cdot \begin{pmatrix} 2/3 \\ 2/3 \\ 1 \\ 1 \end{pmatrix}, \quad t \in \mathbb{R},$$

as above. Notice again, that this solution is the sum of a particular solution of our initial system and the general solution of the corresponding homogeneous system  $Ax = 0$ , cf. the Kronecker-Capelli Theorem in 2.3.5. We also mention that in the Gauss elimination method, free variables are the variables whose column in the reduced row echelon form does not contain any pivot (e.g.,  $x_4$  is the free variable of this specific example, and indeed, the fourth column in REEF is the unique column which does not contain any pivot).

The reduced echelon form of a given matrix  $B$  can be quickly described in Sage, by the command `B.rref()`. Thus, the following cell verifies the (necessarily unique) expression presented above.

```
A=matrix([[3, 0, 3, -5], [1, -1, 1, -1],
[-2, -1, 4, -2], [2, 1, -1, -1]])
b=vector([-8, -2, 0, -3])
B=A.augment(b)
show(B.rref())
```

□

**2.A.15. Remarks on Sage methods.** Although we noticed the importance of `B.echelon_form()` in 2.A.10 and 2.A.12 (mimicking our handling of matrices by hand), the Sage method `B.rref()` seems to be much more efficient. We should also mention that Sage provides methods for directly identifying the pivots of the augmented matrix. The relevant command is `B.pivots()` (or we can use the command `B.pivot_rows()`).

For example, adding in the previous cell `B.pivots()`, Sage answers  $(0, 1, 2)$ , which means that the pivots are in the first, second and third columns.

Recall now that a linear system of equations is *inconsistent* (does not have a solution), if there is a pivot in the last column of an echelon form of the augmented matrix. Therefore, we can combine this with the previous commands in Sage to derive quickly if a linear system of equations is consistent or not.

**2.A.16.** Use Sage to find the reduced echelon form and the pivot entries of the following matrices:

$$B_{\pm} = \begin{pmatrix} 1 & 1 & -2 \\ 3 & 2 & 0 \\ 2 & \pm 1 & 2 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & 1 & 7 & 0 \\ -1 & 4 & 10 & 5 \\ 3 & 2 & 12 & 10 \\ 0 & 0 & 0 & 10 \end{pmatrix}.$$

Next, deduce about the consistency of

- (a) the linear systems having as extended matrix one of  $B_{\pm}$ ;
- (b) the homogeneous linear system  $Cx = 0$ . ○

rank of both matrices is the same, then the backwards elimination provides exactly as many free parameters as the difference between the number of variables  $n$  and the rank  $h(A)$ . In particular, there will be exactly one solution if and only if the matrix is invertible.

All this will be stated explicitly in terms of abstract vector spaces in the important Kronecker-Capelli theorem, see 2.3.5.

## 2. Determinants



In the fifth part of the first chapter, we introduced the scalar function  $\det$  on square matrices of dimension 2 over the real numbers, called determinant, see 1.5.5. We saw that the determinant assigned a non-zero number to a matrix if and only if the matrix was invertible. We did not say it in exactly this way, but you can check for yourself in previous paragraphs starting with 1.5.4 and formula 1.5.5(1).

We saw also that determinants were useful in another way, see the paragraphs 1.5.10 and 1.5.11. There we showed that the area of the parallelepiped should be linearly dependent on every two of the vectors defining it. It was useful to require the change of the sign when changing the order of these vectors. Because determinants (and only determinants) have these properties, up to a constant scalar multiple, we concluded that it was determining the area. Now we will see that we can proceed similarly for every finite dimension.

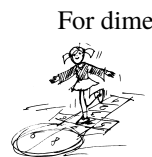
We work again with arbitrary scalars  $\mathbb{K}$  and matrices over these scalars. Our results about determinants will thus hold for all commutative rings, notably also for integer matrices or matrices over any residue classes.

**2.2.1. Definition of the determinant.** Each bijective mapping from a set  $X$  to itself is called a *permutation of the set*  $X$ , cf. 1.3.1. If  $X = \{1, 2, \dots, n\}$ , the permutation can be written by putting the resulting ordering into a table:

$$\begin{pmatrix} 1 & 2 & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(n) \end{pmatrix}.$$


In this way, we shall view a permutation as a bijection or an ordering.

The element  $x \in X$  is called a fixed point of the permutation  $\sigma$  if  $\sigma(x) = x$ . If there exist exactly two distinct elements  $x, y \in X$  such that  $\sigma(x) = y$  while all other elements  $z \in X$  are fixed points, then the permutation  $\sigma$  is called a *transposition*, and we denote it by  $(x, y)$ . Of course, then  $\sigma(y) = x$  holds for such a transformation.



For dimension 2, the formula for a determinant was simple – take all possible products of two elements, one from every column and every row of the matrix, give them a sign such that interchanging two columns leads to the change of the sign of the whole result, and sum all of them (that is, both):

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \det A = ad - bc.$$

 Our next focus is on computing inverses  $A^{-1}$  of matrices  $A \in \text{Mat}_n(\mathbb{K})$  for fields  $\mathbb{K}$ . Viewing  $X = A^{-1}$  as an unknown matrix, we aim at  $AX = E$  (the identity matrix). This provides  $n$  systems of linear equations with the same coefficient matrix  $A$  (we view the matrix equation column-wise).

We already touched this in the case  $n = 2$ . We realize that the general case is equivalent to finding RREF for the augmented matrix  $(A \mid E)$ .

**2.A.17.** Solve the following matrix equations:

$$\begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix} X_1 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad X_2 \begin{pmatrix} 1 & 3 \\ 3 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

**2.A.18.** Compute the inverse of the matrices



$$A = \begin{pmatrix} 4 & 3 & 2 \\ 5 & 6 & 3 \\ 3 & 5 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 1 \\ 3 & 3 & 4 \\ 2 & 2 & 3 \end{pmatrix}.$$

Then determine the matrix  $(A^T \cdot B)^{-1}$ .

**Solution.** As just explained, an  $n \times n$  matrix  $A$  is invertible if and only if  $A$  is row-equivalent to the  $n \times n$  identity matrix  $E$ . Hence, to find the inverse we first form the augmented matrix  $(A \mid E)$ , which in our case reads as

$$(A \mid E) = \left( \begin{array}{ccc|ccc} 4 & 3 & 2 & 1 & 0 & 0 \\ 5 & 6 & 3 & 0 & 1 & 0 \\ 3 & 5 & 2 & 0 & 0 & 1 \end{array} \right).$$

The goal now is to transform the submatrix  $A$  of the augmented matrix into the identity matrix using elementary row operations on  $(A \mid E)$ . This process illustrates how Gaussian elimination is applied to compute the inverse of a matrix

Note that this transformation may not always be possible, depending on the rank of  $A$ . (Later we shall see that having the maximal rank is equivalent to having the determinant non-zero). However, whenever we can achieve the unit matrix in the place of  $A$ , the original unit submatrix gets transformed into the desired  $A^{-1}$ . Notice, this is exactly solving all the given  $n$  systems simultaneously, using the RREF method with the shared coefficient matrix  $A$ . In this case the final (augmented) matrix will be  $(E \mid A^{-1})$ .

Consider now square matrices  $A = (a_{ij})$  of dimension  $n$  over  $\mathbb{K}$ . The formula for the determinant of the matrix  $A$  is also composed of all possible products from elements from individual rows and columns, with properly chosen signs.

In dimension 3 we can guess the correct signs easily. The product of the elements on the diagonal should be with positive sign and we want anti-symmetry when interchanging two columns or rows. This gives the so called *Sarrus rule*:

SARRUS RULE

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$$

The general definition can be formulated via a sum over all permutations:

DEFINITION OF DETERMINANT

The determinant of the matrix  $A$  is a scalar  $\det A = |A|$  defined by the relation

$$|A| = \sum_{\sigma \in \Sigma_n} \text{sgn}(\sigma) a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where  $\Sigma_n$  is the set of all possible permutations over  $\{1, \dots, n\}$  and the symbol  $\text{sgn}$  for a permutation  $\sigma$ , called the parity of  $\sigma$ , will be described below. Each of the expressions

$$\text{sgn}(\sigma) a_{1\sigma(1)} \cdot a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

is called a *term in the determinant*  $|A|$ .

**2.2.2. Parity of permutation.** How should we define the sign of a permutation? We say that a pair of elements  $a, b \in X = \{1, \dots, n\}$  forms an *inversion in the permutation*  $\sigma$ , if  $a < b$  and  $\sigma(a) > \sigma(b)$ . A permutation  $\sigma$  is called *even* or *odd*, if it contains an even or odd number of inversions, respectively.

Thus, the *parity*  $\text{sgn} \sigma$  of the permutation  $\sigma$  is  $(-1)^{\text{number of inversions}}$  and we denote it by  $\text{sgn}(\sigma)$ . This amounts to our definition of sign for computing determinant. But we should like to know how to calculate the parity. The following theorem reveals that the Sarrus rule really defines the determinant in dimension 3.

**Theorem.** Over the set  $X = \{1, 2, \dots, n\}$  there are exactly  $n!$  distinct permutations. These can be ordered in a sequence such that every two consecutive permutations differ in exactly one transposition. Every transposition changes parity.

For any chosen permutation  $\sigma$  there is such a sequence starting with  $\sigma$ .

**PROOF.** For  $n = 1$  or  $n = 2$ , the claim is trivial. We prove the theorem by induction on the size  $n$  of the set  $X$ .

Assume that the claim holds for all sets with  $n - 1$  elements and consider a permutation  $\sigma(1) = a_1, \dots, \sigma(n) =$

We compute

$$\begin{aligned}
 (A \mid E) \quad R_1 \rightarrow \widetilde{R_1 - R_3} & \begin{pmatrix} 1 & -2 & 0 & \mid & 1 & 0 & -1 \\ 5 & 6 & 3 & \mid & 0 & 1 & 0 \\ 3 & 5 & 2 & \mid & 0 & 0 & 1 \end{pmatrix} \\
 R_2 \rightarrow \widetilde{R_2 - 5R_1} \\
 R_3 \rightarrow \widetilde{R_3 - 3R_1} & \begin{pmatrix} 1 & -2 & 0 & \mid & 1 & 0 & -1 \\ 0 & 16 & 3 & \mid & -5 & 1 & 5 \\ 0 & 11 & 2 & \mid & -3 & 0 & 4 \end{pmatrix} \\
 R_2 \rightarrow \widetilde{R_2 - R_3} & \begin{pmatrix} 1 & -2 & 0 & \mid & 1 & 0 & -1 \\ 0 & 5 & 1 & \mid & -2 & 1 & 1 \\ 0 & 11 & 2 & \mid & -3 & 0 & 4 \end{pmatrix} \\
 R_3 \rightarrow \widetilde{R_3 - 2R_2} & \begin{pmatrix} 1 & -2 & 0 & \mid & 1 & 0 & -1 \\ 0 & 5 & 1 & \mid & -2 & 1 & 1 \\ 0 & 1 & 0 & \mid & 1 & -2 & 2 \end{pmatrix} \\
 R_1 \rightarrow \widetilde{R_1 + 2R_3} \\
 R_2 \rightarrow \widetilde{R_2 - 5R_3} & \begin{pmatrix} 1 & 0 & 0 & \mid & 3 & -4 & 3 \\ 0 & 0 & 1 & \mid & -7 & 11 & -9 \\ 0 & 1 & 0 & \mid & 1 & -2 & 2 \end{pmatrix} \\
 R_2 \leftrightarrow R_3 & \begin{pmatrix} 1 & 0 & 0 & \mid & 3 & -4 & 3 \\ 0 & 1 & 0 & \mid & 1 & -2 & 2 \\ 0 & 0 & 1 & \mid & -7 & 11 & -9 \end{pmatrix},
 \end{aligned}$$

where the row operations performed at each step are indicated for clarity. This means that

$$A^{-1} = \begin{pmatrix} 3 & -4 & 3 \\ 1 & -2 & 2 \\ -7 & 11 & -9 \end{pmatrix}.$$

As a verification it is easy to see that  $AA^{-1} = A^{-1}A = E$ , or we can directly use the command `A.inverse()` in Sage:

```
A=matrix([[4, 3, 2],[5, 6, 3],[3, 5, 2]])
show(A); show(A.inverse())
```

Observe now that when calculating  $A^{-1}$ , we did not have to cope with fractions thanks to the suitably chosen row transformations (and the fact, that  $A^{-1}$  is an integer matrix, too!). We leave the similar steps leading to  $B^{-1}$  to the reader (again, the result happens to be an integer matrix):

$$B^{-1} = \begin{pmatrix} 1 & 2 & -3 \\ -1 & 1 & -1 \\ 0 & -2 & 3 \end{pmatrix}.$$

Based now on the identity

$$(A^T \cdot B)^{-1} = B^{-1} \cdot (A^T)^{-1} = B^{-1} \cdot (A^{-1})^T$$

we finally obtain  $(A^T \cdot B)^{-1} = \begin{pmatrix} -14 & -9 & 42 \\ -10 & -5 & 27 \\ 17 & 10 & -49 \end{pmatrix}.$

To describe this result in Sage one can continue typing in the previous cell, by adding the code

```
B=matrix([[1, 0, 1],[3, 3, 4],[2, 2, 3]])
show(B.inverse())
show((transpose(A)*B).inverse())
```

$a_n$ . According to the induction assumption, all the permutations that end with  $a_n$  can be obtained in a sequence, where every two consecutive permutations differ in one transposition. There are  $(n-1)!$  such permutations. In order to proceed further, we select the last of them, and use the transposition of  $\sigma(n) = a_n$  with some element  $a_i$  which has not been at the last position yet. Once again, we form a sequence of all permutations that end with  $a_i$ . After doing this procedure  $n$ -times, we obtain  $n(n-1)! = n!$  distinct permutations – that is, all permutations on  $n$  elements. The resulting sequence satisfies the condition.

Note that the last sentence of the theorem does not seem to be useful in practice. But it is a very important part for proving the theorem by induction over the size of  $X$ .

It remains to prove the part of the theorem about parities. Consider the ordering

$$(a_1, \dots, a_i, a_{i+1}, \dots, a_n),$$

containing  $r$  inversions. Then in the ordering

$$(a_1, \dots, a_{i+1}, a_i, \dots, a_n)$$

there are either  $r-1$  or  $r+1$  inversions. Every transposition  $(a_i, a_j)$  is obtainable by doing  $(j-i) + (j-i-1) = 2(j-i)-1$  transpositions of neighbouring elements.

Therefore any transposition changes the parity. Also, we already know that all permutations can be obtained by applying transpositions.  $\square$

We found that applying a transposition changes the parity of a permutation and any ordering of numbers  $\{1, 2, \dots, n\}$  can be obtained through transposing of neighbouring elements. Therefore we have proven

**Corollary.** *On every finite set  $X = \{1, \dots, n\}$  with  $n$  elements,  $n > 1$ , there are exactly  $\frac{1}{2}n!$  even permutations, and  $\frac{1}{2}n!$  odd permutations.*

If we compose two permutations, it means first doing all transpositions forming the first permutation and then all the transpositions forming the second one. Therefore for any two permutations  $\sigma, \eta : X \rightarrow X$  we have

$$\text{sgn}(\sigma \circ \eta) = \text{sgn}(\sigma) \cdot \text{sgn}(\eta)$$

and also

$$\text{sgn}(\sigma^{-1}) = \text{sgn}(\sigma).$$

**2.2.3. Decomposing permutations into cycles.** A good tool for practical work with permutations is the cycle decomposition, which is also a good exercise on the concept of equivalence.

which also verifies the given expression for  $B^{-1}$ . □

**2.A.19.** Compute the inverse of the following matrices and next provide a verification by Sage:

$$A = \begin{pmatrix} 1 & 0 & -2 \\ 2 & -2 & 1 \\ 5 & -5 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 8 & 3 & 0 & 0 & 0 \\ 5 & 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 & 5 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & i \\ -i & 3 \end{pmatrix},$$

where as usual  $i = \sqrt{-1}$ . ○

*We have seen that a system of linear equations  $Ax = b$  may have: (1) no solution; (2) a unique solution; (3) infinitely many solutions, depending on one or more free parameters.*

*In the homogeneous case, where  $b = 0$ , there is always (at least) one solution. The collections of all solutions forms a vector space (computed as the kernel of the mapping  $x \mapsto Ax$ , we shall come back to these concepts in more abstract way later).*

*If the right-hand side vector  $b$  has at least one non-zero entry, we say that the system is non-homogeneous. Then, the space of all solutions is closely related to the ranks of the matrices  $A$  and  $(A | b)$ .*

*The rank of a matrix is defined as the maximal number of its linearly independent rows (which is the same as the maximal number of linearly independent columns in  $A$ , cf. 2.1.11). This corresponds to the fact that solutions  $x$  of the system  $Ax = b$  are the coefficients in the expression of  $b$  as a linear combination of the columns of  $A$ . We have seen that the case (1) happens if the rank of the augmented matrix is bigger than that of  $A$  (which of course cannot happen if  $b = 0$ ). The number of free parameters for the cases (2) and (3) is the difference between the size of  $b$  and the rank of  $A$ . Later we will see that the rank is closely related to the nullity of the matrix, see for example the description in ??.*

**2.A.20.** Determine the rank of the matrix

$$A = \begin{pmatrix} 1 & -3 & 0 & 1 \\ 1 & -2 & 2 & -4 \\ 1 & -1 & 0 & 1 \\ -2 & -1 & 1 & -2 \end{pmatrix}.$$

Then, determine the number of solutions of the system

$$\begin{cases} x_1 + x_2 + x_3 - 2x_4 = 4, \\ -3x_1 - 2x_2 - x_3 - x_4 = 5, \\ \quad + 2x_2 \quad \quad + x_4 = 1, \\ x_1 - 4x_2 + x_3 - 2x_4 = 3. \end{cases}$$

CYCLES

A permutation  $\sigma$  over the set  $X = \{1, \dots, n\}$  is called a cycle of length  $k$ , if we can find elements  $a_1, \dots, a_k \in X$ ,  $2 \leq k \leq n$  such that  $\sigma(a_i) = a_{i+1}$ ,  $i = 1, \dots, k - 1$ , while  $\sigma(a_k) = a_1$ , and other elements in  $X$  are fixed-points of  $\sigma$ . Cycles of length two are transpositions.

**Proposition.** *Every permutation is a composition of cycles. Cycles of even length have parity  $-1$ , cycles of odd length have parity  $1$ .*

**PROOF.** Fix a permutation  $\sigma$  and define a relation  $R$  such that two elements  $x, y \in X$  are  $R$ -related if and only if  $\sigma^\ell(x) = y$  for some iteration  $\ell \in \mathbb{Z}$  of the permutation  $\sigma$  (notice  $\sigma^{-1}$  means the inverse bijection to  $\sigma$ ). Clearly, it is an equivalence relation (check it carefully!). Because  $X$  is a finite set, for some  $\ell$  it must be that  $\sigma^\ell(x) = x$ . If we pick one equivalence class  $\{x, \sigma(x), \dots, \sigma^{\ell-1}(x)\} \subset X$  and define other elements to be fixed-points, we obtain a cycle. Evidently, the original permutation  $X$  is then the composition of all these cycles for individual equivalence classes and it does not matter in which order we compose the cycles.

For determining the parity we just have to note that cycles of even length can be written as a composition of an odd number of transposition, therefore their parity is  $-1$ . Analogously, cycle of odd length can be obtained using an even number of transpositions and therefore it has parity  $1$ . □

**2.2.4. Expansion of determinant.** Our understanding of the permutations allows to find the *expansion* method of computing the determinants. The simple idea is to collect the terms containing an element in a fixed row in the determinant sum and to add these contributions along the row.

Consider a matrix  $A = (a_{ij})$  and let us look at all terms in  $|A|$  containing the element  $a_{11}$ . By the very definition, these terms correspond to all permutations  $\sigma$  with  $\sigma(1) = 1$ . Thus, the contribution of all these terms to  $|A|$  is  $a_{11}A_{11}$ , where  $A_{11}$  is the determinant of the matrix obtained from  $A$  by omitting the first row and the first column.

Similarly, we can take any other fixed element  $a_{ij}$  in  $A$  and look for the contribution of all terms containing it. Again, we could write  $A_{ij}$  for the determinant of the matrix obtained from  $A$  by omitting the  $i$ -th row and the  $j$ -th column, and the latter contribution must have terms like in  $a_{ij}A_{ij}$ , but we have to be very careful about the signs. While the actual terms of  $|A|$  would be  $\text{sgn } \sigma a_{i_j} a_{1_{\sigma(1)}} \dots \hat{\dots} a_{n_{\sigma(n)}}$  where the hat denotes the omission of the  $i$ -th entry and  $\sigma(i) = j$ , the signatures of the permutations in  $A_{ij}$ , with the  $i$  and  $j$  omitted might be different.

In order to compare it to the previous case  $i = 1, j = 1$ , we can change the initial ordering of the elements in the domain and target of the permutations  $\sigma$ . Clearly,  $i - 1$  changes on the domain and  $j - 1$  changes on the target do the job



Moreover, find all solutions of the corresponding homogeneous system, and in addition of the system

$$\begin{cases} x_1 - 3x_2 & = & 1, \\ x_1 - 2x_2 + 2x_3 & = & -4, \\ x_1 - x_2 & = & 1, \\ -2x_1 - x_2 + x_3 & = & -2. \end{cases} \quad \circ$$

**B. Determinants**

A key object in matrix calculus is the scalar function  $\det$ , the so called determinant, which we already came across when discussing the dimension two calculus (see for example 1.E.25, 1.E.29). Now we shall handle the determinant concept for general square matrices of size  $n$ .

In dimension two, the formula was  $\det A = \det(a_{ij}) = a_{11}a_{22} - a_{12}a_{21}$ , i.e., each term picked up just one entry from each row and column and there are just two options there. Again, we want to see a function linear in all individual columns and changing the sign, whenever two columns are exchanged. Thus, we want to include all terms with any permutation of columns, to be chosen for the individual rows.

This leads us back to the permutations, i.e., bijections  $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ , see the 3rd section of Chapter 1, and 2.2.1. In particular, by 2.2.1 we know that a convenient way to denote permutations is the so called two-row notation (a table form for the function  $\sigma$ ). This notation is adopted in this column, as well.

**2.B.1.** Decompose the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 1 & 6 & 7 & 8 & 9 & 5 & 4 & 2 \end{pmatrix}$$

into a product of cycles and product of transpositions.

**Solution.** The term “cycle” is closely related to invariant subsets in  $\{1, \dots, n\}$  under iterated actions of  $\sigma$ . We may easily see them from the defining table: start with the first element and look on the second row, iterating  $\sigma$ . We see that 1 is mapped to 3, while the second iteration sends 3 to 6, and so on. We continue until we again reach the starting element (which must happen since the set is finite). We obtain the following sequence of elements, which map to each other under the given permutation:

$$1 \mapsto 3 \mapsto 6 \mapsto 9 \mapsto 2 \mapsto 1.$$

Such a sequence is called a *cycle*, or better we extend it to a bijection by letting all the other elements as fixpoints (see 2.2.3), and we may denote it by  $(1, 3, 6, 9, 2)$ . Next, choose any element which belongs to the set  $\{1, 2, \dots, 9\}$  but does not belong to our cycle  $(1, 3, 6, 9, 2)$ , e.g., the integer 4. Applying the same procedure as before, we obtain the cycle  $(4, 7, 5, 8)$ . Observe that each element from the set  $\{1, 2, \dots, 9\}$  appears now in one of the previous two cycles. Thus, we can express  $\sigma$  as the composition of these two cycles, i.e.,

$$\sigma = (1, 3, 6, 9, 2) \circ (4, 7, 5, 8) = (4, 7, 5, 8) \circ (1, 3, 6, 9, 2),$$

(by “bubbling” the index in question to the first position by consecutive swaps of neighboring positions).

Thus, the sign correction is  $(-1)^{i+j-2}$  and we have to adjust the value of  $A_{ij}$  as in the following algorithm, which is the simplest version of the more general Laplace expansion formula, see 2.2.9 below. The readers not sure about the details of our argumentation here may wait for the detailed proof in the more general situation.

EXPANSION OF DETERMINANT

The algebraic complement  $A_{ij}$  of the element  $a_{ij}$  in a matrix  $A$  is the  $(-1)^{i+j}$ -multiple of the determinant of the matrix obtained from  $A$  by omitting the  $i$ -th row and the  $j$ -th column.

Fixing the  $i$ -th row or  $j$ -th column,

$$|A| = \sum_{j=1}^n a_{ij}A_{ij}, \quad |A| = \sum_{i=1}^n a_{ij}A_{ij}.$$

The latter formulae correspond to splitting the determinant sum to parts containing terms with the individual elements in the row or column.

For example, an easy application derives the Sarrus rule from the formula in dimension 2 now.

**2.2.5. Simple properties.** Knowing the properties of permutations and their parities from previous paragraphs allows us to derive quickly basic properties of determinants.

For every matrix  $A = (a_{ij})$  of the type  $m/n$  over scalars from  $\mathbb{K}$  we define the *transpose* of  $A$  as the matrix  $A^T = (a'_{ij})$  with elements  $a'_{ij} = a_{ji}$ . The matrix  $A^T$  is of the type  $n/m$ .

A square matrix  $A$  with the property  $A = A^T$  is called *symmetric*. If  $A = -A^T$ , then  $A$  is called *antisymmetric*.

SIMPLE PROPERTIES OF DETERMINANTS

**Theorem.** Every square matrix  $A = (a_{ij})$  satisfies the following conditions:

- (1)  $|A^T| = |A|$ .
- (2) If one of the rows contains only zero elements from  $\mathbb{K}$ , then  $|A| = 0$ .
- (3) If a matrix  $B$  was obtained from  $A$  by transposing two rows, then  $|A| = -|B|$ .
- (4) If a matrix  $B$  was obtained from  $A$  by multiplying one row by a scalar  $a \in \mathbb{K}$ , then  $|B| = a|A|$ .
- (5) If all elements of the  $k$ -th row in  $A$  are of the form  $a_{kj} = c_{kj} + b_{kj}$  and all remaining rows in the matrices  $A, B = (b_{ij}), C = (c_{ij})$  are identical, then  $|A| = |B| + |C|$ .
- (6) A determinant  $|A|$  does not change if we add to any row of  $A$  a linear combination of other rows.

where the second equality occurs since independent cycles commute.

Next, we notice how to decompose general cycles into the simplest ones - the transpositions  $(i, j)$  (i.e.,  $i \mapsto j \mapsto i$  and all other remain fixed). A simple check reveals (do it carefully yourselves!) that any cycle  $(i_1, i_2, \dots, i_k)$  is a composition of transpositions of the neighbouring numbers, starting from the back. Thus, in our case

$$\begin{aligned} (1, 3, 6, 9, 2) &= (1, 3) \circ (3, 6) \circ (6, 9) \circ (9, 2) \\ &\equiv (1, 3)(3, 6)(6, 9)(9, 2). \end{aligned}$$

For example, evaluating this resulting bijection on 2:

$$\begin{aligned} ((1, 3)(3, 6)(6, 9)(9, 2))(2) &= ((1, 3)(3, 6)(6, 9))(9) = \\ &= ((1, 3)(3, 6))(6) = (1, 3)(3) = 1, \end{aligned}$$

as expected.

This means  $\sigma = (1, 3)(3, 6)(6, 9)(9, 2)(4, 7)(7, 5)(5, 8)$  (notice that this is also the decomposition into the minimal number of transpositions).

As we saw in Chapter 1, Sage offers a rich portfolio of methods available for solving combinatorial tasks. To treat the given task in Sage, we need to define a permutation  $p$  of a finite set, which can be done via the command `Permutation(p)`. Also, we can determine the cycles of  $p$  by the command `Permutation(p).to_cycles()`. Let us illustrate the situation for  $p = \sigma$ , where  $\sigma$  is as above:

```
p = Permutation([3, 1, 6, 7, 8, 9, 5, 4, 2])
c = (p.to_cycles())
ts = [] # collecting transpositions
for x in c:
    for i in range(len(x) - 1):
        ts.append( (x[i], x[i + 1]) )
print(ts)
```

This block returns the following list verifying our answer:

```
[(1, 3), (3, 6), (6, 9), (9, 2), (4, 7),
(7, 5), (5, 8)]
```

We are ready to summarize: The right formula for the determinant of a matrix  $A = (a_{ij})$  of size  $n$  is

$$\det(A) = |A| = \sum_{\sigma \in \Sigma_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)},$$

where the parity  $\operatorname{sgn}(\sigma)$  is  $\pm 1$ , according to the even or odd number of permutations constituting  $\sigma$ . This is well defined. Clearly this definition fulfills our wish to have an expression linear in each of the columns, while changing its sign with each transposition of columns.

Note that the minimal number of transpositions in the decomposition of a given permutation can be obtained as follows: We first decompose the permutation into a product of independent cycles, and then the cycles canonically into the transpositions (actually the length of the cycles tells their parity and they are multiplied).

You may find more details in 2.2.2, where the parity is defined in the terms of the so called inversions. In the two-row

PROOF. (1) The terms of determinants  $|A|$  and  $|A^T|$  are in bijective correspondence, where the term  $\operatorname{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}$  corresponds the following  $A^T$  term (notice, multiplication does not depend on the order of scalars)



$$\begin{aligned} \operatorname{sgn}(\sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n} &= \\ &= \operatorname{sgn}(\sigma) a_{1\sigma^{-1}(1)} \cdots a_{n\sigma^{-1}(n)}, \end{aligned}$$

and we have to ensure that this member has the correct sign. But the parities of  $\sigma$  and  $\sigma^{-1}$  are the same, and so this is really a term in the determinant  $|A^T|$  and the first claim is proved.

(2) This comes straight from the definition of determinant, because all its terms contain exactly one member from every row. Thus, if one of the rows is zero, all terms of the determinant are also zero.

(3) The only change in the terms of  $|B|$  compared to  $|A|$  is the addition of one transposition in all permutations, therefore all the signs will be reversed.

(4) This follows straight from the definition, because terms of  $|B|$  are just terms of  $|A|$  multiplied by the scalar  $a$ .

(5) In every term of  $|A|$ , there is exactly one element from the  $k$ -th row of the matrix  $A$ . By the distributive law for multiplication and addition in  $\mathbb{K}$ , the claim follows directly from the definition of determinant.

(6) If there are two identical rows in  $A$ , then there are always two identical terms among all terms in the determinant, up to the sign. Therefore in this case  $|A| = 0$ . Thus, by (5), we can add any other row to the given row, without changing the value of the determinant. In view of the claims (4) and (5), we can in fact add a scalar multiple of any other row.  $\square$

### 2.2.6. Computational corollaries.

Let us note a nice corollary of the first claim of the previous theorem about the equality of the determinants of the matrix and its transpose. It ensures that whenever we prove some claim about determinants formulated in terms of rows of the corresponding matrix, we immediately obtain an analogous claim in terms of the columns.



For instance, we can immediately formulate all the claims (2)–(6) for linear combinations of columns.

For a while, let us assume, we work with a field of scalars  $\mathbb{K}$ . Then, by the previous theorem, we can use elementary row transformations to bring any square matrix  $A$  into row echelon form, without changing the value of its determinant. We just have to be careful and add only linear combinations of other rows to a given one.

Thus let us look at the distribution of the elements in the individual terms of a determinant  $|A|$  with dimension of  $A$  equal to  $n > 1$ . There is just one term with all of its elements on the diagonal. In all other terms, there must be elements both above and below the diagonal (if we place one element outside of the diagonal, we block two diagonal entries and we leave only  $n - 2$  diagonal positions for the other  $n - 1$  elements).

notation for permutations, the number of inversions is seen when you go through the columns and count the number of those where the second row is smaller than the first one.

**2.B.2.** Determine the parity of the permutation  $\sigma$  in 2.B.1, and of the permutation  $\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 6 & 1 & 5 & 3 \end{pmatrix}$ .  $\circ$

**2.B.3.** Check the previous result in Sage.

**Solution.** It is very easy to obtain the parity of a permutation in Sage. This is done via the command `p.sign()`, where  $p$  is the given permutation. For instance, for the given  $\sigma$  and  $\tau$  in 2.B.2 we can type:

```
s=Permutation([3,1,6,7,8,9,5,4,2])
t=Permutation([2,4,6,1,5,3])
print(s.sign()); print(t.sign())
```

Sage says  $\sigma, \tau$  are odd by returning  $-1$  for both cases.  $\square$

**2.B.4.** Compute the determinant of the matrices

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 & 3 \\ 1 & -1 & 2 \\ 3 & 2 & 2 \end{pmatrix}, C = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

**Solution.** For  $A$  we compute  $\det(A) = 1 \cdot 1 - 2 \cdot 2 = -3$ . As for the  $3 \times 3$  matrices, the definition yields the so called Sarrus rule, see 2.2.1. In particular, for  $B$  we get

$$\det(B) = 1 \cdot (-1) \cdot 2 + 2 \cdot 2 \cdot 3 + 3 \cdot 1 \cdot 2 - 3 \cdot (-1) \cdot 3 - 1 \cdot 2 \cdot 2 - 1 \cdot 2 \cdot 2 = 17.$$

Another way to remember this rule:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ \ell & m & n \end{pmatrix} = a \cdot \alpha - b \cdot \beta - c \cdot \gamma,$$

where  $\alpha = \det \begin{pmatrix} e & f \\ m & n \end{pmatrix}, \beta = \det \begin{pmatrix} d & f \\ \ell & n \end{pmatrix}, \gamma = \det \begin{pmatrix} d & e \\ \ell & m \end{pmatrix}$ .

Similarly, for the matrix  $C$  we obtain

$$\det(C) = 1 \cdot 0 \cdot 1 + 1 \cdot 0 \cdot 1 + 1 \cdot 0 \cdot (-2) - 1 \cdot 0 \cdot (-2) - 1 \cdot 1 \cdot 1 - 1 \cdot 0 \cdot 0 = -1.$$

Of course, all cases can be done in Sage, as in Chapter 1. Let us recall the relevant code:

```
B=matrix([[1,2,3],[1,-1,2],[3,2,2]])
det(B)
```

There are several methods for computing determinants of big matrices. One of them is to use the elementary row (or column) transformations. Remember though that we are allowed to add linear combinations of other rows to the transformed one, since multiplying a row clearly results in the same multiple of the resulting determinant. If  $A$  is in a row-echelon (column-echelon) form, then  $\det(A)$  is just the multiple of all diagonal elements (straight from definition!). See a more detailed review of elementary properties in 2.2.5.



Therefore, if the matrix  $A$  is in a row echelon form, then every term of  $|A|$  is zero, except the term with exclusively diagonal entries. This proves the following algorithm:

COMPUTING DETERMINANTS USING ELIMINATION

**Lemma.** If  $A$  is in the row echelon form then

$$|A| = a_{11}a_{22} \cdots a_{nn}.$$

This observation gives an effective method for computing determinants using the Gauss elimination method for matrices over a field of scalars  $\mathbb{K}$ , see the paragraph 2.1.7.

Notice that the very same argumentation allows us to stop the elimination having the first  $k$  columns in the requested form and finding the determinant of the matrix  $B$  of dimension  $n-k$  in the right bottom corner of  $A$  in another way. The result will then be  $|A| = a_{11}a_{22} \cdots a_{kk}|B|$ .

As a useful (theoretical) illustration of this principle, we shall derive the following formula for direct calculation of solutions of systems of linear equations. For sake of simplicity, we still work with field of scalars now. (But we shall see later, that Cramer rule actually works for all scalars.)



CRAMER RULE

**Proposition.** Consider the system of  $n$  linear equations for  $n$  variables with matrix of the system  $A = (a_{ij})$  and the column of values  $b = (b_1, \dots, b_n)$ . In matrix notation this means we are solving the equation  $A \cdot x = b$ .

If there exists the inverse  $|A|^{-1}$ , then the individual components of the unique solution  $x = (x_1, \dots, x_n)$  are given as

$$x_i = |A_i||A|^{-1},$$

where the matrices  $A_i$  arise from the matrix  $A$  of the system by replacing the  $i$ -th column by the column  $b$  of values.

**PROOF.** Even for general scalars, if  $A^{-1}$  exists, then there must be a unique solution. As we have already seen, working over field of scalars the inverse of the matrix of the system exists if and only if the system has a unique solution, and this in turn happens if and only if  $|A| \neq 0$ .

If we have such a solution  $x$ , we can express the column  $b$  in the matrix  $A_i$  by the corresponding linear combination of the columns of the matrix  $A$ , that is the values  $b_k = a_{k1}x_1 + \cdots + a_{kn}x_n$ . Then, by subtracting the  $x_\ell$ -multiples of all the other  $\ell$ -th columns from this  $i$ -th column in  $A_i$ , we arrive at just the  $x_i$ -multiple of the original column of  $A$ . The number  $x_i$  can thus be brought in front of the determinant to obtain the equation  $|A_i| = x_i |A|$ , and thus  $|A_i||A|^{-1} = x_i|A||A|^{-1} = x_i$ , which is our claim.  $\square$

Notice also that the properties (3)–(5) from the previous theorem say that the determinant, (considered as a mapping which assigns a scalar to  $n$  vectors of dimension  $n$ ), is an

The other option is to use the Laplace expansion with respect to one or more rows or columns, see the description in 2.2.9.

**2.B.5.** Compute  $\det(B)$  for the matrix  $B$  from the previous task by the Gauss elimination method.

**Solution.** We aim at transforming the matrix into a row echelon form, and then multiplying the numbers on the diagonal. However, we must remember that a multiplication of a row with a scalar changes the determinant of the matrix by the same multiple. Moreover, interchanging two rows changes the sign of the determinant of the matrix.

$$\begin{vmatrix} 1 & 2 & 3 \\ 1 & -1 & 2 \\ 3 & 2 & 2 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 \\ 0 & -3 & -1 \\ 0 & -4 & -7 \end{vmatrix} = \frac{1}{-4} \cdot \frac{1}{3} \cdot \begin{vmatrix} 1 & 2 & 3 \\ 0 & 12 & 4 \\ 0 & -12 & -21 \end{vmatrix} \\ = -\frac{1}{12} \cdot \begin{vmatrix} 1 & 2 & 3 \\ 0 & 12 & 4 \\ 0 & 0 & -17 \end{vmatrix}.$$

The remaining step is the computation of the determinant of an upper triangular matrix. The latter equals to the product of the diagonal entries, thus,  $\det(B) = -\frac{1}{12}(1 \cdot 12 \cdot (-17)) = 17$ . Of course, direct application of the Sarrus rule was faster.  $\square$

**2.B.6.** Compute the determinant of the matrix

$$A = \begin{pmatrix} 1 & 3 & 5 & 6 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{pmatrix}.$$

**Solution.** A standard way to compute  $\det(A)$  is the reduction of the size of the matrices by the expansion along the first column (or line, see also 2.3.10). This gives

$$\det(A) = 1 \cdot \begin{vmatrix} 2 & 2 & 2 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{vmatrix} - 1 \cdot \begin{vmatrix} 3 & 5 & 6 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \end{vmatrix} + 1 \cdot \begin{vmatrix} 3 & 5 & 6 \\ 2 & 2 & 2 \\ 1 & 2 & 1 \end{vmatrix} \\ = -2 - 2 + 6 = 2,$$

where we have exploited that the last entry of the first column is zero.

An alternative way is to convert the matrix to row echelon form, which we can do as follows:

$$\det(A) = - \begin{vmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 5 & 6 \\ 0 & 1 & 2 & 1 \end{vmatrix} = - \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 2 & 4 & 4 \\ 0 & 1 & 2 & 1 \end{vmatrix} = \\ = - \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 1 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 4 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{vmatrix} = 2.$$

Observe that here we have interchanged the rows twice. Specify the rest of the applied rules, according to 2.2.5!  $\square$

**2.B.7.** Use the Laplace theorem (see 2.2.9) to compute the determinant of the matrix  $D$  given below. Next verify your

antisymmetric mapping linear in every argument, exactly as we required in analogy to the 2-dimensional case.

**2.2.7. Further properties of the determinant.** Later we



will see that, exactly as in the dimension 2, the determinant of the matrix equals to the (oriented) volume of the parallelepiped determined by the columns of the matrix. We shall also see that considering the mapping  $x \mapsto A \cdot x$  given by the square matrix  $A$  on  $\mathbb{R}^n$  we can understand the determinant of this matrix as expressing the ratio between the volume of the parallelepipeds given by the vectors  $x_1, \dots, x_n$  and their images  $A \cdot x_1, \dots, A \cdot x_n$ .

Because the composition  $x \mapsto A \cdot x \mapsto B \cdot (A \cdot x)$  of mappings corresponds to the matrix multiplication, the Cauchy theorem below is easy to understand:

CAUCHY THEOREM

**Theorem.** Let  $A = (a_{ij}), B = (b_{ij})$  be square matrices of dimension  $n$  over the ring of scalars  $\mathbb{K}$ . Then

$$|A \cdot B| = |A| \cdot |B|.$$

In the next paragraphs, we derive this theorem in a purely algebraic way, in particular because the previous argumentation based on geometrical intuition could hardly work for arbitrary scalars. The basic tool is the *determinant expansion* using one or more of the rows or columns which we have seen in simplest case of single rows or columns in 2.2.4.

We will also need a little technical preparation. The reader who is not fond of too much abstraction can skip these paragraphs and note only the statement of the Laplace theorem and its corollaries.

Notice also, the claims (2), (3) and (6) from the theorem 2.2.5 are easily deduced from the Cauchy theorem and the representation of the elementary row transformations as multiplication by suitable matrices (cf. 2.1.8).

**2.2.8. Minors of the matrix.** When investigating matrices



and their properties we often work only with parts of the matrices. Therefore we need some new concepts.

SUBMATRICES AND MINORS

Let  $A = (a_{ij})$  be a matrix of the type  $m/n$  and let  $1 \leq i_1 < \dots < i_k \leq m, 1 \leq j_1 < \dots < j_\ell \leq n$  be fixed natural numbers. Then the matrix

$$M = \begin{pmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \dots & a_{i_1 j_\ell} \\ \vdots & & & \vdots \\ a_{i_k j_1} & a_{i_k j_2} & \dots & a_{i_k j_\ell} \end{pmatrix}$$

of the type  $k/\ell$  is called a *submatrix of the matrix A* determined by the rows  $i_1, \dots, i_k$  and columns  $j_1, \dots, j_\ell$ . The remaining  $(m - k)$  rows and  $(n - \ell)$  columns determine a matrix  $M^*$  of the type  $(m - k)/(n - \ell)$ , which is called *complementary submatrix* to  $M$  in  $A$ .

answer in Sage.

$$D = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 3 & 0 & 3 \\ 4 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}. \quad \circ$$

**2.B.8.** This is a theoretical task. Let us fix any matrices  $A, B \in \text{Mat}_n(\mathbb{K})$ , over any ring  $\mathbb{K}$ .

(a) If  $\det(A) = 0$ , show that the product  $AB$  also has determinant zero.

(b) Prove the identity  $\det(cA) = c^n \det(A)$ , for each scalar  $c \in \mathbb{K}$ .

(c) Show by a counterexample that in general  $\det(A + B) \neq \det(A) + \det(B)$ . ○

The upcoming series of exercises will explore various methods for computing the inverse of an invertible matrix, including Cramer's rule, and other topics. As we will demonstrate, Sage can efficiently handle these computations, leveraging its extensive matrix calculus functionality.

Before diving into the problems, let us review some key points about matrix manipulation in Sage. There are several convenient shortcuts for matrix creation. As we pronounced, to specify the ring of the matrix entries use `QQ` for rationals, `RR` for reals, `CC` for complex numbers, and `ZZ` for integers. For exact computations with rational entries the results are precise. For numerical calculations with real or complex matrices, we recommend to use of `RDF` and `CDF` working with floating-point representation, instead of `RR` and `CC`.

After creating a matrix  $A$  in your favourite Sage editor, you can access its entries in a straightforward manner using the square bracket notation `A[i, j]`. Here,  $i$  and  $j$ , represent the row and column indices, respectively. Note that both Python and Sage use zero-based indexing.

For example, to access the entry  $a_{3,3}$  of a  $4 \times 4$  matrix  $A$ , we type `A[2, 2]`. Notice, Sage also allows indexing from the end of the matrix, using negative indices. For example, `A[-1, :]`, refers to the last row of  $A$ , and `A[-2, :]` refers to the second to last row of  $A$ . Additionally, indices can be specified as intervals using the format `a : b`. For instance, `A[2 : 3]` is valid and will be useful when discussing submatrices later. Another useful tip is the notation `p : q : r`, which generates all indices from  $p$  to  $q - 1$ , in steps of  $r$ .

**2.B.9.** Given the matrix  $A = \begin{pmatrix} 1 & 1 & 3 & 0 & 8 \\ 2 & 2 & -4 & -8 & 4 \\ -1 & -5 & 2 & 4 & 6 \\ 7 & -1 & 10 & 0 & 9 \end{pmatrix}$  use Sage to access the entries  $a_{33}$ ,  $a_{12}$ ,  $a_{11}$ ,  $a_{45}$  and  $a_{44}$ .

**Solution.** We can use the cell

```
A=matrix(QQ, [[1, 1, 3, 0, 8], [2, 2, -4, -8, 4],
[-1, -5, 2, 4, 6], [7, -1, 10, 0, 9]])
print(A[2, 2]); print(A[0, 1])
print(A[0, 0]); print(A[3, 4])
print(A[3, 3])
```

When  $k = \ell$  we call the determinant  $|M|$  the *subdeterminant* or *minor* of the order  $k$  of the matrix  $A$ . If  $m = n$  and  $k = \ell$ , then  $M^*$  is also a square matrix and  $|M^*|$  is called the *minor complement* to  $|M|$ , or *complementary minor* of the submatrix  $M$  in the matrix  $A$ . The scalar

$$(-1)^{i_1 + \dots + i_k + j_1 + \dots + j_l} \cdot |M^*|$$

is then called the *algebraic complement* of the minor  $|M|$ .

The submatrices formed by the first  $k$  rows and columns are called *leading principal submatrices*, and their determinants are called *leading principal minors* of the matrix  $A$ . If we choose  $k$  sequential rows and columns starting with the  $i$ -th row, we speak of *principal submatrices* and *principal minors*.

Specially, when  $k = \ell = 1$ ,  $m = n$  we call the corresponding algebraic complementary minor the *algebraic complement*  $A_{ij}$  of the element  $a_{ij}$  of the matrix  $A$ , which we met already in 2.2.4.

**2.2.9. Laplace determinant expansion.** If the leading principal minor  $|M|$  of the matrix  $A$  is of the order  $k$ , then, directly from the definition of the determinant, each of the individual  $k!(n - k)!$  terms in the product of  $|M|$  with its algebraic complement is a term of  $|A|$ .

In general, consider a square submatrix  $M$ , that is, a square matrix given by the rows  $i_1 < i_2 < \dots < i_k$  and columns  $j_1 < \dots < j_k$ . Then using  $(i_1 - 1) + \dots + (i_k - k)$  exchanges of neighbouring rows and  $(j_1 - 1) + \dots + (j_k - k)$  exchanges of neighbouring columns in  $A$  we can transform this submatrix  $M$  into a leading principal submatrix and the complementary matrix gets transformed into its complementary matrix.

The whole matrix  $A$  gets transformed into a matrix  $B$  satisfying (cf. 2.2.5 and the definition of the determinant)  $|B| = (-1)^\alpha |A|$ , where  $\alpha = \sum_{h=1}^k (i_h + j_h) - 2(1 + \dots + k)$ . But  $(-1)^\alpha = (-1)^\beta$  with  $\beta = \sum_{h=1}^k (i_h + j_h)$ . Therefore we have checked:

**Proposition.** If  $A$  is a square matrix of dimension  $n$  and  $|M|$  is its minor of the order  $k < n$ , then the product of any term of  $|M|$  with any term of its algebraic complement is a term in the determinant  $|A|$ .

This claim suggests that we could perhaps express the determinant of the matrix by using some products of smaller determinants. We see that  $|A|$  contains exactly  $n!$  distinct terms, exactly one for each permutation. These terms are mutually distinct as polynomials in the components of a general matrix  $A$ . If we can show that there are exactly that many mutually distinct expressions from the previous claim, we obtain the determinant  $|A|$  as their sum.

It remains to show that the terms of the product  $|M| \cdot |M^*|$  contain exactly  $n!$  distinct members from  $|A|$ .

From the chosen  $k$  rows we can choose  $\binom{n}{k}$  minors  $M$  and using the previous lemma each of the  $k!(n - k)!$

Sage returns 2, 1, 1, 9 and 0, respectively. Notice entering the command  $A([4, 3])$ , Sage will produce an error message: matrix index out of range, which is expected.  $\square$

**2.B.10.** For the matrices  $A, B$  given below, compute:

- (a) the corresponding null spaces;
- (b) the submatrix of  $A$  obtained by canceling the first row, and the first and third columns; Is this diagonal?
- (c) the submatrix of  $B$  obtained by canceling the first and last row, and the last column.
- (d) the submatrix of  $B$  obtained by canceling the first and third row, and the second and fourth column. Is this symmetric?

$$A = \begin{pmatrix} 1 & -1 & \pi & 0 \\ 0 & \pi & -1 & 0 \\ 1 & 0 & -1 & \pi \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{pmatrix}.$$

**2.B.11.** Using the Cramer’s rule solve the system posed in 2.A.9. Next demonstrate the situation in Sage.

**Solution.** Consider a linear system  $Ax = b$  of  $n$  equations and  $n$  unknowns, with  $\det(A) \neq 0$ . Recall by 2.2.6 that the Cramer’s rule describes the unique solution of such a linear system as the unique vector  $x = (x_1, \dots, x_n)^T$  of  $\mathbb{R}^n$  with entries  $x_i = \frac{\det(A_i)}{\det(A)}$ , for all  $i = 1, \dots, n$ , where  $A_i$  are the matrixes obtained from the coefficient matrix  $A$  by replacing the  $i$ th column by the vector  $b$ , for  $i = 1, \dots, n$ . The matrix  $A$  of the system in 2.A.9 is  $3 \times 3$  and the Cramer’s rule gives

$$x_1 = \frac{\begin{vmatrix} 2 & 2 & 3 \\ -3 & -3 & -1 \\ -3 & 1 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & 1 \\ -3 & 1 & 2 \end{vmatrix}} = 1, \quad x_2 = \frac{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & -1 \\ -3 & -3 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & 1 \\ -3 & 1 & 2 \end{vmatrix}} = 2,$$

$$x_3 = \frac{\begin{vmatrix} 1 & 2 & 2 \\ 2 & -3 & -3 \\ -3 & 1 & -3 \end{vmatrix}}{\begin{vmatrix} 1 & 2 & 3 \\ 2 & -3 & 1 \\ -3 & 1 & 2 \end{vmatrix}} = -1.$$

It is easy to demonstrate Cramer’s rule in Sage. For a  $3 \times 3$  system this can done really quickly, and an illustration is here:

```
A=matrix([[1, 2, 3], [2, -3, -1], [-3, 1, 2]])
b=vector([2, -3, -3])
A1=copy(A); A2=copy(A); A3=copy(A)
A1[:,0] = b #construct the matrix A_1
A2[:,1] = b #construct the matrix A_2
A3[:,2] = b #construct the matrix A_3
show(A1); show(A2); show(A3)
x1=A1.det()/A.det(); x2=A2.det()/A.det()
x3=A3.det()/A.det(); print([x1, x2, x3])
```

terms in the products of  $|M|$  with their algebraic complements is a term in  $|A|$ . But for distinct choices of  $M$  we can never obtain the same terms and the individual terms in  $(-1)^{i_1+\dots+i_k+j_1+\dots+j_l} \cdot |M| \cdot |M^*|$  are also mutually distinct. Therefore we have exactly the required number  $k!(n-k)!\binom{n}{k} = n!$  of terms, and we have proved:

LAPLACE THEOREM

**Theorem.** Let  $A = (a_{ij})$  be a square matrix of dimension  $n$  over arbitrary ring of scalars with  $k$  rows fixed. Then  $|A|$  is a sum of all  $\binom{n}{k}$  products  $(-1)^{i_1+\dots+i_k+j_1+\dots+j_l} \cdot |M| \cdot |M^*|$  of minors of the order  $k$  chosen among the fixed rows with their algebraic complements.

The Laplace theorem transforms the computation of  $|A|$  into the computation of determinants of lower dimension. This method of computation is called the *Laplace expansion* along the chosen rows (or columns). For instance, the expansion along the  $i$ -th row or the  $j$ -th column is:

$$|A| = \sum_{j=1}^n a_{ij} A_{ij} = \sum_{i=1}^n a_{ij} A_{ij}$$

where  $A_{ij}$  are the algebraic complements of the elements  $a_{ij}$  (that is, minors of order one), as deduced in 2.2.4 already.

In practical computations, it is often efficient to combine the Laplace expansion with a direct method of Gaussian elimination.

**2.2.10. Proof of the Cauchy theorem.** The theorem is based on a clever but elementary application of the Laplace theorem. We just use the Laplace expansion twice on a particular arrangement of a well chosen matrix.

Consider first the following matrix  $H$  of dimension  $2n$  (we are using the so-called block symbolics, that is, we write the matrix as if composed of the (sub)matrixes  $A, B$ , and so on).

$$H = \begin{pmatrix} A & 0 \\ -E & B \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} & 0 & \dots & 0 \\ -1 & & 0 & b_{11} & \dots & b_{1n} \\ & \ddots & & \vdots & & \vdots \\ 0 & & -1 & b_{n1} & \dots & b_{nn} \end{pmatrix}$$

The Laplace expansion along the first  $n$  rows gives

$$|H| = |A| \cdot |B|.$$

Now in sequence, we add linear combinations of the first  $n$  columns to the last  $n$  columns in order to obtain a matrix

Notice, the `copy()` function provides a way to make a copy of a matrix before we make changes to it. This block returns the matrices  $A_1, A_2, A_3$  and the solution in the form  $[1, 2, -1]$ . In Section E we will meet another implementation of Cramer's rule via Sage, see 2.E.23.  $\square$

**2.B.12.** Is the matrix  $A$  given below invertible? In the positive case find its algebraically adjoint matrix and its inverse:

$$A = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 3 & 0 & 4 \\ 5 & 0 & 6 & 0 \\ 0 & 7 & 0 & 8 \end{pmatrix}.$$

**Solution.** Recall that a square matrix  $A$  is invertible if and only if  $\det(A) \neq 0$ . Expanding the first row of the given  $A$ , we see that

$$\det(A) = \begin{vmatrix} 1 & 0 & 2 & 0 \\ 0 & 3 & 0 & 4 \\ 5 & 0 & 6 & 0 \\ 0 & 7 & 0 & 8 \end{vmatrix} = \begin{vmatrix} 3 & 0 & 4 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} + 2 \begin{vmatrix} 0 & 3 & 4 \\ 5 & 0 & 0 \\ 0 & 7 & 8 \end{vmatrix} = 16.$$

Hence there exists  $A^{-1}$  such that  $AA^{-1} = E = A^{-1}A$ , where  $E$  is the  $4 \times 4$  identity matrix. To compute  $A^{-1}$  we will use the algebraic adjoint matrix of  $A$ , the latter given by

$$\text{adj}(A) = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}^T.$$

Here,  $A_{ij}$  is the algebraic complement of the element  $a_{ij}$  of the matrix  $A$ , that is, the product of the number  $(-1)^{i+j}$  and the determinant of the matrix given by  $A$  without the  $i$ th row and  $j$ th column. We compute:

$$\begin{aligned} A_{11} &= \begin{vmatrix} 3 & 0 & 4 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} = -24, & A_{12} &= - \begin{vmatrix} 0 & 0 & 4 \\ 5 & 6 & 0 \\ 0 & 0 & 8 \end{vmatrix} = 0, \\ A_{13} &= \begin{vmatrix} 0 & 3 & 4 \\ 5 & 0 & 0 \\ 0 & 7 & 8 \end{vmatrix} = 20, & A_{14} &= - \begin{vmatrix} 0 & 3 & 0 \\ 5 & 0 & 6 \\ 0 & 7 & 0 \end{vmatrix} = 0, \\ A_{21} &= - \begin{vmatrix} 0 & 2 & 0 \\ 0 & 6 & 0 \\ 7 & 0 & 8 \end{vmatrix} = 0, & A_{22} &= \begin{vmatrix} 1 & 2 & 0 \\ 5 & 6 & 0 \\ 0 & 0 & 8 \end{vmatrix} = -32, \\ A_{23} &= - \begin{vmatrix} 1 & 0 & 0 \\ 5 & 0 & 0 \\ 0 & 7 & 8 \end{vmatrix} = 0, & A_{24} &= \begin{vmatrix} 1 & 0 & 2 \\ 5 & 0 & 6 \\ 0 & 7 & 0 \end{vmatrix} = -28, \\ A_{31} &= \begin{vmatrix} 0 & 2 & 0 \\ 3 & 0 & 4 \\ 7 & 0 & 8 \end{vmatrix} = 8, & A_{32} &= - \begin{vmatrix} 1 & 2 & 0 \\ 0 & 0 & 4 \\ 0 & 0 & 8 \end{vmatrix} = -0, \\ A_{33} &= \begin{vmatrix} 1 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 7 & 8 \end{vmatrix} = -4, & A_{34} &= - \begin{vmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 0 & 7 & 0 \end{vmatrix} = -0, \\ A_{41} &= - \begin{vmatrix} 0 & 2 & 0 \\ 3 & 0 & 4 \\ 0 & 6 & 0 \end{vmatrix} = 0, & A_{42} &= \begin{vmatrix} 1 & 2 & 0 \\ 0 & 0 & 4 \\ 5 & 6 & 0 \end{vmatrix} = -16, \end{aligned}$$

with zeros in the bottom right corner. We obtain

$$K = \begin{pmatrix} a_{11} & \cdots & a_{1n} & c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} & c_{n1} & \cdots & c_{nn} \\ -1 & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & -1 & 0 & \cdots & 0 \end{pmatrix}.$$

The elements of the submatrix on the top right part must satisfy

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj},$$

that is, they are exactly the components of the product  $A \cdot B$  and  $|K| = |H|$ . The expansion of the last  $n$  columns gives us  $|K| = (-1)^n (-1)^{1+\cdots+2n} |A \cdot B| = (-1)^{2n \cdot (n+1)} \cdot |A \cdot B| = |A \cdot B|$ . This proves the Cauchy theorem.

**2.2.11. Determinant and the inverse matrix.** Assume first



that there is an inverse matrix of the matrix  $A$ , that is,  $A \cdot A^{-1} = E$ . Since the unit matrix always satisfies  $|E| = 1$ , it follows that for every invertible matrix its determinant is an invertible scalar and by the Cauchy theorem we have  $|A^{-1}| = |A|^{-1}$ .

But we can say more, combining the Laplace and Cauchy theorems.

#### INVERSE MATRIX DETERMINANT FORMULA

For any square matrix  $A = (a_{ij})$  of dimension  $n$  we define the matrix  $A^* = (a_{ij}^*)$ , where  $a_{ij}^* = A_{ji}$  are algebraic complements of the elements  $a_{ji}$  in  $A$ . The matrix  $A^*$  is called the *algebraically adjoint matrix* of the matrix  $A$  (or the *adjugate matrix*  $A^*$ ).

**Theorem.** For every square matrix  $A$  over a ring of scalars  $\mathbb{K}$  we have that

$$(1) \quad AA^* = A^*A = |A| \cdot E.$$

In particular,

- (i)  $A^{-1}$  exists as a matrix over the ring of scalars  $\mathbb{K}$  if and only if  $|A|^{-1}$  exists in  $\mathbb{K}$ .
- (ii) If  $A^{-1}$  exists, then  $A^{-1} = |A|^{-1} \cdot A^*$ .



**PROOF.** As already mentioned, the Cauchy theorem shows that the existence of  $A^{-1}$  implies the invertibility of  $|A| \in \mathbb{K}$ .

For an arbitrary square matrix  $A$  we can directly compute  $A \cdot A^* = (c_{ij})$ , where

$$c_{ij} = \sum_{k=1}^n a_{ik} a_{kj}^* = \sum_{k=1}^n a_{ik} A_{jk}.$$

If  $i = j$ , it is exactly the Laplace expansion of  $|A|$  along the  $i$ -th row.

If  $i \neq j$ , then we may imagine we expand the determinant along the  $j$ -th row, but plug in the values of the  $i$ -th row instead of the  $a_{jk}$ 's. This is the expansion of the determinant

$$A_{43} = - \begin{vmatrix} 1 & 0 & 0 \\ 0 & 3 & 4 \\ 5 & 0 & 0 \end{vmatrix} = 0, \quad A_{44} = \begin{vmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 5 & 0 & 6 \end{vmatrix} = -12.$$

Thus, by substitution we obtain that

$$\begin{aligned} \text{adj}(A) &= \begin{pmatrix} -24 & 0 & 20 & 0 \\ 0 & -32 & 0 & 28 \\ 8 & 0 & -4 & 0 \\ 0 & 16 & 0 & -12 \end{pmatrix}^T \\ &= \begin{pmatrix} -24 & 0 & 8 & 0 \\ 0 & -32 & 0 & 16 \\ 20 & 0 & -4 & 0 \\ 0 & 28 & 0 & -12 \end{pmatrix}. \end{aligned}$$

As a verification of the given expression in Sage, you can use either the command `A.adjugate`, or its alias `A.adjoint_classical()`, as follows:

```
A=matrix([[1, 0, 2, 0],[0, 3, 0, 4],
[5, 0, 6, 0], [0, 7, 0, 8]])
show(A.adjugate())
```

The inverse matrix  $A^{-1}$  is now obtained by the rule  $A^{-1} = \frac{1}{\det(A)} \cdot \text{adj}(A)$ , that is,

$$A^{-1} = \begin{pmatrix} -3/2 & 0 & 1/2 & 0 \\ 0 & -2 & 0 & 1 \\ 5/4 & 0 & -1/4 & 0 \\ 0 & 7/4 & 0 & -3/4 \end{pmatrix}.$$

Recall that we can directly compute either the determinant of  $A$ , or the inverse of  $A$  in Sage by adding the commands `det(A)` and `A.inverse()`, respectively.  $\square$

### C. Vector spaces and linear mappings

*Intuitively, a vector space is a landscape (of any dimension  $n$ , finite or infinite) where each position corresponds to a unique set of scalar coordinates (again  $n$ ) and we add and multiply them exactly as we already saw in chapter 1, working with  $\mathbb{R}^2$  and  $\mathbb{C}$ .*

*Vector spaces can become quite abstract and unusual (as already the scalars can be) and they are fundamental building blocks in mathematics, with an amazing wide range of applications in many fields, including chemistry, physics, image processing, computer science, economics, etc. For an informative introduction to the concepts, see 2.3.1. Next we will primarily focus on finite-dimensional vector spaces.*

**2.C.1. Vector spaces.** Decide whether the following sets form a vector space over the given field:

(a) The set of real solutions of the linear system:

$$\begin{cases} x_1 + x_2 + \cdots + x_8 + x_9 + x_{10} & = & 10x_1, \\ x_1 + x_2 + \cdots + x_8 + x_9 & = & 9x_1, \\ x_1 + x_2 + \cdots + x_8 & = & 8x_1, \\ \vdots & & \\ x_1 + x_2 & = & 2x_1. \end{cases}$$

of a matrix where the  $i$ -th and  $j$ -th row is the same, therefore  $c_{ij} = 0$ .

This implies that  $A \cdot A^* = |A| \cdot E$ , and we have proven one of the equalities (1). In particular, if  $|A|^{-1}$  exists, then  $A \cdot (|A|^{-1}A^*) = E$ .

If  $|A|$  is an invertible scalar, we may repeat the previous computation for  $A^* \cdot A$ , and we obtain  $(|A|^{-1}A^*) \cdot A = E$ . Therefore our computation really gives the inverse matrix of  $A$ , as claimed in the theorem.  $\square$

Notice that for fields of scalars we have already proved that the right inverse of a matrix is automatically the left inverse and thus the inverse, too. Here we have obtained the same result for all rings of scalars, together with a strong and effective existence condition. On the other hand the exact formula for the inverse has become rather theoretical with little practical value.

As a direct corollary of this theorem we can once again prove the Cramer rule for solving the systems of linear equations, see 2.2.6. Really, for the solution of the system  $A \cdot x = b$  we just need to read in the equation

$$x = A^{-1} \cdot b = |A|^{-1}A^* \cdot b$$

the individual components of the expression  $A^* \cdot b$  as the Laplace expansions of the determinant of the matrix  $A_i$  which arose through the exchange of the  $i$ -th column of  $A$  for the column  $b$ .

## 3. Vector spaces and linear mappings

**2.3.1. Abstract vector spaces.** Let us go back for a while to the systems of  $m$  linear equations of  $n$  variables from 2.1.3 and further, let us assume that the system is homogeneous,  $A \cdot x = 0$ , i.e.,



$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

By the distributivity of the matrix multiplication it is clear that the sum of two solutions  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$  satisfies

$$A \cdot (x + y) = A \cdot x + A \cdot y = 0$$

and thus is also a solution. Similarly, a scalar multiple  $a \cdot x$  is also a solution. The set of all solutions of a fixed system of equations is therefore closed under vector addition and scalar multiplication. These are the basic properties of vectors of dimension  $n$  in  $\mathbb{K}^n$ , see 2.1.1. Now we have the vectors in the solution space with  $n$  coordinates. The “dimension” of this space is given by the difference of the number of variables and the rank of the matrix  $A$ . Thus we can easily deal with the solution of a system of 1000 equations in 1000 variables and need only one or two free parameters. Thus the whole solution space will behave as a plane or a line, as we have already seen in 1.5.3 at the page 29, although the vectors themselves are given by so many components.



(b) The set of real solutions of the equation

$$x_1 + x_2 + \cdots + x_{10} = 0.$$

(c) The set of real solutions of the equation

$$x_1 + 2x_2 + 3x_3 + \cdots + 10x_{10} = 1.$$

(d) The set of all real or complex sequences (recall that a real or a complex sequence can be viewed as a mapping  $f : \mathbb{N} \rightarrow \mathbb{R}$  or  $f : \mathbb{N} \rightarrow \mathbb{C}$ , respectively).

(e) The set  $\mathcal{F} = \{f \mid f : X \rightarrow \mathbb{R}\}$  of real-valued functions with domain a non-empty set  $X$ .

(f) The set of real solutions of a non-homogeneous difference equation.

(g) The set  $\text{Mat}_{m,n}(\mathbb{K})$  of  $m \times n$  matrices with entries over  $\mathbb{K}$ , where  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ .

(h) The set  $\mathbb{Q}(\sqrt{2}) = \{a + \sqrt{2}b : a, b \in \mathbb{Q}\}$  over  $\mathbb{Q}$ .

**Solution.** (a) This set is a vector space, since the solutions are real multiples of the vector  $u := (1, \dots, 1)^T$  (10-factors). Moreover, a sum of two multiples of the same vector is again a multiple of the vector, while the reverse vector is again a multiple of the vector. All other axioms are trivially satisfied.

(b) This provides an example of a vector space of dimension 9 (its dimension is determined by the number of free parameters of the solution). More general, recall that the set of solutions of any homogeneous linear system is a vector space, see 2.3.5.

(c) Taking twice the solution  $x_1 = 1, x_i = 0$ , for  $i = 2, \dots, 10$ , we do not obtain a solution. Hence, the given set cannot be a vector space. However, the set of solutions forms an affine space, see 4.1.1.

(d) The set of all real (respectively, complex) sequences clearly forms a real (respectively, complex) vector space. Notice that the addition and scalar multiplication is defined term-wise. Moreover, the zero vector is represented by the zero sequence  $(0, 0, 0, \dots)$ .

(e) This provides an example of an infinite-dimensional vector space, where the addition and multiplication are defined by the usual way, i.e.,  $(f + g)(x) = f(x) + g(x)$  and  $(cf)(x) = c \cdot f(x)$ , respectively, for all  $x \in X$ . Notice the zero vector is simply the constant function with zero value everywhere on  $X$ , i.e., the zero function defined on  $X$ .

(f) Consider two solutions of a non-homogeneous equation

$$\begin{aligned} a_n x_{n+k} + a_{n-1} x_{n+k-1} + \cdots + a_0 x_k &= c, \\ a_n y_{n+k} + a_{n-1} y_{n+k-1} + \cdots + a_0 y_k &= c, \end{aligned}$$

where  $c$  is some non-zero scalar. Their sum satisfies

$$\begin{aligned} a_n(x_{n+k} + y_{n+k}) + a_{n-1}(x_{n+k-1} + y_{n+k-1}) \\ + \cdots + a_0(x_k + y_k) = 2c, \end{aligned}$$

and obviously, this cannot provide a solution of the given non-homogeneous equation. Hence we do not obtain a vector space. However, the set of solutions in this case forms an affine space (see 4.1.1).

(g) This is one of the most fundamental examples of finite-dimensional vector spaces. For simplicity, let us fix  $\mathbb{K} = \mathbb{R}$  and similarly is treated the complex case, see also the task

We go further. Already in paragraph 1.2.1 we have encountered an interesting example of a space of all solutions of a homogeneous linear difference equation of first order. All solutions have been obtained from a single one by scalar multiplication and are also closed under addition and scalar multiples. These “vectors” of solutions are infinite sequences of numbers, although we intuitively expect that the “dimension” of the whole space of solutions should be one. We shall understand such phenomena with the help of a more general definition of vector space and its dimension.

#### VECTOR SPACE DEFINITION

A vector space  $V$  over a field of scalars  $\mathbb{K}$  is a set where we define the operations

- addition, which satisfies the axioms (CG1)–(CG4) from the paragraph 1.1.1 on the page 5,
- scalar multiplication, for which the axioms (V1)–(V4) from the paragraph 2.1.1 on the page 84 hold.

Recall our simple notational convention: scalars are usually denoted by letters from the beginning of the alphabet, that is,  $a, b, c, \dots$ , while for vectors we shall use letters from the end, that is,  $u, v, w, x, y, z$ . Usually,  $x, y, z$  will denote  $n$ -tuples of scalars. For completeness, the letters from the centre of the alphabet, for instance  $i, j, k, \ell$ , will mostly denote indices.

In order to gain some practice in the formal approach, we check some simple properties of vectors. These are trivial for  $n$ -tuples for scalars, but not so evident for general vectors in our new abstract sense.



**2.3.2. Proposition.** Let  $V$  be a vector space over a field of scalars  $\mathbb{K}$ . Suppose  $a, b, a_i \in \mathbb{K}$ , and  $u, v, u_j \in V$ . Then

- (1)  $a \cdot u = 0$  if and only if  $a = 0$  or  $u = 0$ ,
- (2)  $(-1) \cdot u = -u$ ,
- (3)  $a \cdot (u - v) = a \cdot u - a \cdot v$ ,
- (4)  $(a - b) \cdot u = a \cdot u - b \cdot u$ ,
- (5)  $(\sum_{i=1}^n a_i) \cdot (\sum_{j=1}^m u_j) = \sum_{i=1}^n \sum_{j=1}^m a_i \cdot u_j$ .

**PROOF.** We can expand

$$a \cdot u = (a + 0) \cdot u \stackrel{(V2)}{=} a \cdot u + 0 \cdot u$$

which, according to the axiom (CG4), implies  $0 \cdot u = 0$ . Now

$$u + (-1) \cdot u \stackrel{(V2)}{=} (1 + (-1)) \cdot u = 0 \cdot u = 0$$

and thus  $-u = (-1) \cdot u$ . Further,

$$a \cdot (u + (-1) \cdot v) \stackrel{(V2, V3)}{=} a \cdot u + (-a) \cdot v = a \cdot u - a \cdot v,$$

which proves (3). It follows that

$$(a - b) \cdot u \stackrel{(V2, V3)}{=} a \cdot u + (-b) \cdot u = a \cdot u - b \cdot u$$

which proves (4). Property (5) follows using induction with (V2) and (V1).

It remains to prove (1):  $a \cdot 0 = a \cdot (u - u) = a \cdot u - a \cdot u = 0$ , which along with the first derived proposition in

in 2.E.56 for more details on complex matrices. The vector addition on  $\text{Mat}_{m,n}(\mathbb{R})$  is given by matrix addition, i.e.,  $A + B = (a_{ij} + b_{ij}) \in \text{Mat}_{m,n}(\mathbb{R})$  for all  $A = (a_{ij})$ ,  $B = (b_{ij}) \in \text{Mat}_{m,n}(\mathbb{R})$ , and scalar multiplication is defined by multiplying each entry of the given matrix by the scalar, i.e.,  $cA = (c \cdot a_{ij}) \in \text{Mat}_{m,n}(\mathbb{R})$  for all  $c \in \mathbb{R}$  and  $A = (a_{ij}) \in \text{Mat}_{m,n}(\mathbb{R})$ , see also 2.1.2. Let us check the axioms (CG1)-(CG4) presented in 1.1.1:

- (CG1) – associativity: Indeed,  $(A+B)+C = A+(B+C)$  for all  $A, B, C \in \text{Mat}_{m,n}(\mathbb{R})$ . This property follows by the associativity of the addition of real numbers:  $(a_{ij} + b_{ij}) + c_{ij} = a_{ij} + (b_{ij} + c_{ij})$ , where we assume that  $A = (a_{ij})$ ,  $B = (b_{ij})$  and  $C = (c_{ij})$ , respectively.
  - (CG2) – commutativity: In an analogous way we can prove that  $A + B = B + A$ , for all  $A, B \in \text{Mat}_{m,n}(\mathbb{R})$ .
  - (CG3) – existence of neutral element: This is the zero matrix, which we just denote by  $0 \in \text{Mat}_{m,n}(\mathbb{R})$ . Obviously, it satisfies  $A + 0 = A = 0 + A$ , for all  $A \in \text{Mat}_{m,n}(\mathbb{R})$ .
  - (CG4) – existence of additive inverse: For  $A \in \text{Mat}_{m,n}(\mathbb{R})$ ,  $B = -A$  is the unique matrix satisfying  $A + B = 0$ .
- These axioms ensure that the pair  $(\text{Mat}_{m,n}(\mathbb{R}), +)$  has the structure of a “commutative group”.

Finally, it is very easy to check the axioms (V1) – (V4) posed in the paragraph 2.1.1, and which must satisfy the scalar multiplication. For completeness, we list them:

- (V1) –  $\alpha(A + B) = \alpha A + \alpha B$  for all  $\alpha \in \mathbb{R}$ , and  $A, B \in \text{Mat}_{m,n}(\mathbb{R})$ .
- (V2) –  $(\alpha + \beta)A = \alpha A + \beta A$  for all  $\alpha, \beta \in \mathbb{R}$  and  $A \in \text{Mat}_{m,n}(\mathbb{R})$ .
- (V3) –  $\alpha(\beta A) = (\alpha\beta)A$ , for all  $\alpha, \beta \in \mathbb{R}$  and  $A \in \text{Mat}_{m,n}(\mathbb{R})$ .
- (V4) –  $1A = A$  for all  $A \in \text{Mat}_{m,n}(\mathbb{R})$ .

(h) The final set  $\mathbb{Q}(\sqrt{2}) = \{a + \sqrt{2}b : a, b \in \mathbb{Q}\}$  provides an example of a vector space over  $\mathbb{Q}$ . Indeed, let  $x = a + \sqrt{2}b$ ,  $y = c + \sqrt{2}d$  two elements in  $\mathbb{Q}(\sqrt{2})$ . Then  $x + y = (a + c) + \sqrt{2}(b + d) \in \mathbb{Q}(\sqrt{2})$ . Moreover, for  $q \in \mathbb{Q}$  we have  $qx \in \mathbb{Q}(\sqrt{2})$ , since  $qa, qb \in \mathbb{Q}$ , for all  $a, b \in \mathbb{Q}$ .  $\square$

**2.C.2.** By the discussion in 2.3.4 we know that the space  $\mathbb{R}_m[x]$  of real polynomials of degree at most  $m$  is a vector space over  $\mathbb{R}$ . Can we claim the same for the set of real polynomials of degree exactly  $m$ ?  $\circ$

**2.C.3. Vector spaces in Sage.** (a) Use Sage to create the vector spaces  $\mathbb{R}^4$  and  $\mathbb{Q}^4$ . Next check if the vector  $u = 2v + 3w$  is an element of  $\mathbb{Q}^4$ , where  $v = (1, -2, -3, 0)^T$  and  $w = (1, 0, 3, -2)^T$ , respectively.

(b) Use Sage to create the matrix spaces  $\text{Mat}_4(\mathbb{R})$  of  $4 \times 4$  real matrices, and  $\text{Mat}_{4,1}(\mathbb{C})$  of  $4 \times 1$  complex matrices.

**Solution.** (a) An easy way to introduce a vector space in Sage, is to start by specifying the field over which the vectors are defined, and then use an exponent to indicate the dimension of the vector space. This process is compact and mimics the notation we use when working on paper.

For instance, to introduce  $\mathbb{R}^4$ , type

this proof proves one implication. For the other implication, we use an axiom for the field of scalars, and axiom (V4) for vector spaces: if  $p \cdot u = 0$  and  $p \neq 0$ , then  $u = 1 \cdot u = (p^{-1} \cdot p) \cdot u = p^{-1} \cdot 0 = 0$ .  $\square$

**2.3.3. Linear (in)dependence.** In paragraph 2.1.11 we worked with linear combinations of rows of a matrix. With vectors we work analogously:

#### LINEAR COMBINATION AND INDEPENDENCE

An expression of the form  $a_1 v_1 + \dots + a_k v_k$  is called a *linear combination* of vectors  $v_1, \dots, v_k \in V$ .

A finite sequence of vectors  $v_1, \dots, v_k$  is called *linearly independent*, if the only zero linear combination is the one with all coefficients zero. That is, for any scalars  $a_1, \dots, a_k \in \mathbb{K}$ ,  $a_1 v_1 + \dots + a_k v_k = 0$  implies  $a_1 = a_2 = \dots = a_k = 0$ . It is clear that for an independent sequence of vectors, all vectors are mutually distinct and nonzero.

The set of vectors  $M \subset V$  in a vector space  $V$  over  $\mathbb{K}$  is called *linearly independent*, if every finite  $k$ -tuple of vectors  $v_1, \dots, v_k \in M$  is linearly independent.

The set of vectors  $M$  is *linearly dependent*, if it is not linearly independent.



A nonempty subset  $M$  of vectors in a vector space over a field of scalars  $\mathbb{K}$  is dependent if and only if one of its vectors can be expressed as a finite linear combination using other vectors in  $M$ . This follows directly from the definition:

Indeed, at least one of the coefficients in the corresponding linear combination must be nonzero, and since we are over a field of scalars, we can multiply whole combination by the inverse of this nonzero coefficient and thus express its corresponding vector as a linear combination of the others.

Every subset of a linearly independent set  $M$  is clearly also linearly independent (we require the same conditions on a smaller set of vectors). Similarly, we can see that  $M \subset V$  is linearly independent if and only if every finite subset of  $M$  is linearly independent.

**2.3.4. Generators and subspaces.** A subset  $M \subset V$  is called a *vector subspace* if it forms, together with the restricted operations of addition and scalar multiplication, a vector space. That is, we require

$$\forall a, b \in \mathbb{K}, \forall v, w \in M, a \cdot v + b \cdot w \in M.$$

We investigate a couple of cases: The space of  $m$ -tuples of scalars  $\mathbb{R}^m$  with coordinate-wise addition and multiplication is a vector space over  $\mathbb{R}$ , but also a vector space over  $\mathbb{Q}$ . For instance for  $m = 2$ , the vectors  $(1, 0), (0, 1) \in \mathbb{R}^2$  are linearly independent, because from

$$a \cdot (1, 0) + b \cdot (0, 1) = (0, 0)$$

follows  $a = b = 0$ . Further, the vectors  $(1, 0), (\sqrt{2}, 0) \in \mathbb{R}^2$  are linearly dependent over  $\mathbb{R}$ , because  $\sqrt{2} \cdot (1, 0) = (\sqrt{2}, 0)$ , but over  $\mathbb{Q}$  they are linearly independent! Over  $\mathbb{R}$  these two

```
V=RR^4; V # we can also type V=RR**4
```

Sage returns the following:

```
Vector space of dimension 4 over
Real Field with 53 bits of precision
```

Similarly, to introduce the vector space  $\mathbb{Q}^4$  type `QQ^4`, etc.

We can also use the `VectorSpace()` function, which requires the name of the number system for the entries and the number of entries in each vector. For instance the cell

```
V = VectorSpace(QQ, 4); print(V)
```

prints out

```
Vector space of dimension 4
over Rational Field
```

Vectors can be constructed in the usual way, that is,

```
V=QQ^4;
v=vector(QQ, [1, -2, -3, 0])
w=vector(QQ, [1, 0, 3, -2])
```

It is also easy to perform computations with vectors (vector addition and scalar multiplication):<sup>6</sup>

```
V=QQ^4; v=vector(QQ, [1, -2, -3, 0])
w=vector(QQ, [1, 0, 3, -2])
u=2*v+3*w
print(u) # we could type show(u)
u in V
```

Here Sage returns the vector  $u$ , i.e.,  $(5, -4, 3, -6)$  and the word `True`, the latter verifying that  $u \in V$ .

(b) Sage provides built-in functions for working with matrix spaces. This cell creates the space  $\text{Mat}_4(\mathbb{R})$ :

```
M = MatrixSpace(RR, 4); M
```

Similarly, to introduce the space  $\text{Mat}_{4,1}(\mathbb{C})$  use the cell

```
M = MatrixSpace(CC, 4, 1); M
```

Notice, when the number of columns is omitted, it defaults to the number of rows.  $\square$

**2.C.4.** Describe the linear combination  $3v_1 + 2v_2 + 5v_3 + v_4$  of the vectors  $v_1, \dots, v_4$  given below in Sage, using at least two different methods:

$$\begin{aligned} v_1 &= \left(\frac{1}{2}, \frac{2}{3}, 1, 0\right)^T, & v_3 &= \left(0, \frac{1}{5}, \frac{2}{5}, 1\right)^T, \\ v_2 &= (2, 0, 1/7, 0)^T, & v_4 &= (-1, 2, 0, 3)^T. \end{aligned}$$

**2.C.5.** Consider the system  $Ax = b$  given below with a solution given by  $(-2/15, 1/5, 2/15, -11/15)^T$ ,

$$\begin{pmatrix} 3 & -9 & 9 & 0 \\ 9 & -3 & 6 & 0 \\ 9 & -6 & 0 & -6 \\ -9 & 12 & 6 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \end{pmatrix}.$$

<sup>6</sup>Be aware that Sage displays vectors using parentheses (to distinguish them from lists), and presents them horizontally. In particular, in Sage there is no inherent distinction between a “row vector” and a “column vector”. However, when dealing with matrices, it is crucial to make this distinction.

vectors “generate” a one-dimensional subspace, while over  $\mathbb{Q}$  the subspace is “larger”.

Polynomials with real coefficients and of degree at most  $m$  form a vector space  $\mathbb{R}_m[x]$ . We can consider the polynomials as mappings  $f : \mathbb{R} \rightarrow \mathbb{R}$  and define the addition and scalar multiplication like this:  $(f + g)(x) = f(x) + g(x)$ ,  $(a \cdot f)(x) = a \cdot f(x)$ .

Polynomials of all degrees also form a vector space  $\mathbb{R}[x]$  (or  $\mathbb{R}_\infty[x]$ ) and  $\mathbb{R}_m[x] \subset \mathbb{R}_n[x]$  is a vector subspace for any  $m \leq n \leq \infty$ . Further examples of subspaces is given by all even polynomials or all odd polynomials, that is, polynomials satisfying  $f(-x) = \pm f(x)$ .

In complete analogy with polynomials, we can define a vector space structure on a set of all mappings  $\mathbb{R} \rightarrow \mathbb{R}$  or of all mappings  $M \rightarrow V$  of an arbitrary fixed set  $M$  into the vector space  $V$ .

Because the condition in the definition of subspace consists only of universal quantifiers, the intersection of subspaces is still a subspace. We can see this also directly: Let  $W_i, i \in I$ , be vector subspaces in  $V$ ,  $a, b \in \mathbb{K}$ ,  $u, v \in \bigcap_{i \in I} W_i$ . Then  $a \cdot u + b \cdot v \in W_i$  for all  $i \in I$ . Hence  $a \cdot u + b \cdot v \in \bigcap_{i \in I} W_i$ .

It can be noted that the intersection of all subspaces  $W \subset V$  that contain some given set of vectors  $M \subset V$  is a subspace. It is called the *linear span* or *linear hull* of  $M$  and we write  $\text{span } M$ .

We say that a set  $M$  *generates* the subspace  $\text{span } M$ , or that the elements of  $M$  are *generators* of the subspace  $\text{span } M$ .

We formulate a few simple claims about subspace generation:

**Proposition.** For every nonempty set  $M \subset V$ , we have

- (1)  $\text{span } M = \{a_1 \cdot u_1 + \dots + a_k \cdot u_k; k \in \mathbb{N}, a_i \in \mathbb{K}, u_j \in M, j = 1, \dots, k\}$ ;
- (2)  $M = \text{span } M$  if and only if  $M$  is a vector subspace;
- (3) if  $N \subset M$  then  $\text{span } N \subset \text{span } M$  is a vector subspace; the subspace  $\text{span } \emptyset$  generated by the empty subset is the trivial subspace  $\{0\} \subset V$ .

**PROOF.** (1) The set of all linear combinations

$$a_1 u_1 + \dots + a_k u_k$$

on the right-hand side of (1) is clearly a vector subspace and of course it contains  $M$ . On the other hand, each of the linear combinations must be in  $\text{span } M$  and thus the first claim is proved.

Claim (2) follows immediately from claim (1) and from the definition of vector space. Analogously, (1) implies most of the third claim.

Finally, the smallest possible vector subspace is  $\{0\}$ . Notice that the empty set is contained in every subspace and each of them contains the vector  $0$ . This proves the last claim.  $\square$

Use Sage to illustrate that the solution gives us scalars that yield the vector of constants  $b$  as a linear combination of the columns of the coefficient matrix  $A$ . ○

**2.C.6.** Describe the vector space structure of  $\mathbb{Z}_5^3$ . What is the total number of vectors in  $\mathbb{Z}_5^3$ ? Use the cardinality command in Sage to support your answer. ○

Next we will explore problems concerning the concept of linear subspaces. A vector subspace of a vector space  $V$  is a non-empty subset  $W$  of  $V$  which is closed under addition and scalar multiplication. It can be either given as a span of some vectors (i.e., all of their linear combinations), or by some (linear) conditions.<sup>7</sup> Therefore, any subspace of  $V$  is itself a vector space and shares the same fundamental properties and significance, see also the discussion in 2.3.4.

In Sage we have numerous options to explore the properties and characteristics of subspaces. Once we have introduced a vector space  $V$ , a subspace  $W \subset V$  is defined by specifying a set of generators that span the subspace. This can be achieved by the `subspace()` function or with the `span` command.

**2.C.7. Linear subspaces.** Decide whether the following statements are true or false:

- (a) The subset  $U_1 = \{(x, y, z) \in \mathbb{R}^3 : 2x + 7y + z = 1\}$  is a linear subspace of  $\mathbb{R}^3$ .
- (b) The subset  $U_2 = \{(x, y, z) \in \mathbb{R}^3 : 2x^2 + 7y + z = 0\}$  is a linear subspace of  $\mathbb{R}^3$ .
- (c) The subset  $U_3 = \{(x, y, z) \in \mathbb{R}^3 : 2x + 7y + z = 0\}$  is a linear subspace of  $\mathbb{R}^3$ .
- (d) The subset  $U_4 = \{ax^2 + c : a, c \in \mathbb{R}\}$  is a linear subspace of the vector space  $\mathbb{R}_2[x]$  of polynomials with real coefficients and degree at most 2.
- (e) The subset  $U_5 = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : |x_1| = |x_2| = |x_3|\}$  is a linear subspace of  $\mathbb{R}^3$ .
- (f) The subset  $U_6 = \{p \in \mathbb{R}_3[x] : p(1) = 0\}$  is a linear subspace of the vector space  $\mathbb{R}_3[x]$  of polynomials with real coefficients and degree at most 3.
- (g) The set of real solutions of a homogeneous difference equation is a linear subspace of the vector space of all real sequences.
- (h) The set  $\left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} : a, b, c \in \mathbb{R} \right\}$  of  $2 \times 2$  upper triangular matrices is a subspace of  $\text{Mat}_2(\mathbb{R})$ .
- (i) The subset  $V_\lambda = \{x \in \mathbb{R}^n : Ax = \lambda x\}$ , where  $A \in \text{Mat}_n(\mathbb{R})$  and  $\lambda \in \mathbb{R}$ , is a linear subspace of  $\mathbb{R}^n$ .
- (j) The subset  $\{A \in \text{Mat}_n(\mathbb{R}) : A = -A^T\}$  of skew-symmetric  $n \times n$  matrices is a subspace of  $\text{Mat}_n(\mathbb{R})$ .

**Solution.** (a) The subset  $U_1$  cannot be a subspace of  $\mathbb{R}^3$ , since the zero vector does not belong to  $U_1$ .

(b) This is also false, and  $U_2$  cannot be subspace of  $\mathbb{R}^3$ . For example, the vector  $u = (1, 0, -2) \in U_2$ , but  $-u =$

<sup>7</sup>This implies that  $W$  contains the zero vector. Indeed, any  $w \in W$  will have an additive inverse  $-w \in W$  and hence  $0 = w + (-w) \in W$ .

A subset  $M \subset V$  is called a *basis of the vector space*  $V$  if  $\text{span } M = V$  and  $M$  is linearly independent.

A vector space with a finite basis is called *finitely dimensional*. The number of elements of the basis is called the *dimension* of  $V$ .

If  $V$  does not have a finite basis, we say that  $V$  is *infinite dimensional*. We write  $\dim V = k$ ,  $k \in \mathbb{N}$  or  $k = \infty$ .

In order to be satisfied with such a definition of dimension, we must know that different bases of the same space will always have the same number of elements. We shall show this below, cf. 2.3.9. But we note immediately, that the trivial subspace is generated by the empty set, which is an “empty” basis. Thus it has dimension zero.

The linearly independent vectors

$$e_i = (0, \dots, 1, \dots, 0) \in \mathbb{K}^n, \quad i = 1, \dots, n$$

(all zeros, but one value 1 at the  $i$ -th position) are the most useful example of a basis in the vector space  $\mathbb{K}^n$ . We call it the *standard basis* of  $\mathbb{K}^n$ .

**2.3.5. Linear equations again.** It is a good time now to recall the properties of systems of linear equation in terms of abstract vector spaces and their bases. As we have already noted in the introduction to this section (cf. 2.3.1), the set of all solutions of the *homogeneous system*

$$A \cdot x = 0$$

is a vector space. If  $A$  is a matrix with  $m$  rows and  $n$  columns, and the rank of the matrix is  $k$ , then using the row echelon transformation (see 2.1.7) to solve the system, we find that the dimension of the space of all solutions is exactly  $n - k$ .

Indeed, the left hand side of the equation can be understood as the linear combination of the columns of  $A$  with coefficients given by  $x$  and the rank  $k$  of the matrix provides the number of linearly independent columns in  $A$ , thus the dimension of the subspace of all possible linear combinations of the given form. Therefore, after transforming the system into row echelon form, exactly  $m - k$  zero rows remain. In the next step, we are left with exactly  $n - k$  free parameters. By setting one of them to have value one, while all others are zero, we obtain exactly  $n - k$  linearly independent solutions. Then all solutions are given by all the linear combinations of these  $n - k$  solutions. Every  $(n - k)$ -tuple of linearly independent solutions is called a *fundamental system of solutions* of the given homogeneous system of equations. We have proved:

**Proposition.** *The set of all solutions of the homogeneous system of equations  $A \cdot x = 0$ , for  $n$  variables with the matrix  $A$  of rank  $k$ , is a vector subspace in  $\mathbb{K}^n$  of dimension  $n - k$ . Every basis of this space forms a fundamental system of solutions of the given homogeneous system.*

Next, consider the general system of equations  $A \cdot x = b$ . Notice that the columns of the matrix  $A$  are actually images of the vectors of the standard basis in  $\mathbb{K}^n$  under the mapping

$(-1, 0, 2) \notin U_2$ , since  $2 \cdot (-1)^2 + 7 \cdot 0 + 2 = 4 \neq 0$ .

(c) This is true and we leave the verification for practice.

(d) The set  $U_4$  is a linear subspace of  $\mathbb{R}_2[x]$ , since

$$\begin{aligned}(a_1x^2 + c_1) + (a_2x^2 + c_2) &= (a_1 + a_2)x^2 + (c_1 + c_2), \\ k(ax^2 + c) &= (ka)x^2 + kc,\end{aligned}$$

for all real numbers  $a_1, c_1, a_2, c_2, a, c, k \in \mathbb{R}$ . Notice also that for  $a = c = 0$  we obtain the zero polynomial.

(e) The subset  $U_5$  is not a linear subspace of  $\mathbb{R}^3$ , since for example  $(1, 1, 1) + (-1, 1, 1) = (0, 2, 2) \notin U_5$ .

(f) The subset  $U_6$  is clearly a subspace of  $\mathbb{R}_3[x]$ . Indeed, if  $p, q \in U_6$ , and  $c \in \mathbb{R}$ , then we see that

$$\begin{aligned}(p + q)(1) &= p(1) + q(1) = 0, \Rightarrow (p + q) \in U_6, \\ (cp)(1) &= cp(1) = c \cdot 0 = 0, \Rightarrow (cp) \in U_6.\end{aligned}$$

(g) Consider two sequences  $(x_j)_{j=0}^\infty$  and  $(y_j)_{j=0}^\infty$  satisfying the same homogeneous difference equation, that is,

$$\begin{aligned}a_n x_{n+k} + a_{n-1} x_{n+k-1} + \cdots + a_0 x_k &= 0 \\ a_n y_{n+k} + a_{n-1} y_{n+k-1} + \cdots + a_0 y_k &= 0.\end{aligned}$$

By adding these equations, we obtain

$$\begin{aligned}a_n(x_{n+k} + y_{n+k}) + a_{n-1}(x_{n+k-1} + y_{n+k-1}) \\ + \cdots + a_0(x_k + y_k) = 0.\end{aligned}$$

This means that also the sequence  $(x_j + y_j)_{j=0}^\infty$  satisfies the given equation. Analogously, if the sequence  $(x_j)_{j=0}^\infty$  satisfies the given equation, then the same applies for  $(ux_j)_{j=0}^\infty$ , for some  $u \in \mathbb{R}$ . Thus the assertion is true.

(h) This claim is again true, as you can easily verify yourself.

(i) If  $x, y \in V_\lambda$ , then we see that

$$\begin{aligned}A(x + y) &= Ax + Ay = \lambda x + \lambda y = \lambda(x + y), \\ A(cx) &= c(Ax) = c(\lambda x) = \lambda(cx),\end{aligned}$$

and hence  $x + y \in V_\lambda$  and  $cx \in V_\lambda$ , for all  $x, y \in V_\lambda$  and  $c \in \mathbb{R}$ . Hence the claim is true. Non-zero vectors lying on  $V_\lambda$  are called *eigenvectors* of  $A$  corresponding to the *eigenvalue*  $\lambda$ . Thus we have just proved that the *eigenspace*  $V_\lambda$  corresponding to the eigenvalue  $\lambda$  is a subspace of  $\mathbb{R}^n$ . We will learn more about these notions later (see Section 2.D.1). Notice that when  $\lambda = 0$ , then  $V_0 = \text{Ker}(A)$  is the null space (kernel) of the squared matrix  $A$ .

(j) Taking  $A, B \in \text{Mat}_n(\mathbb{R})$  with  $A^T = -A$  and  $B^T = -B$ , we see that  $(A + B)^T = -(A + B)$  and  $(cA)^T = -(cA)$ , for all  $c \in \mathbb{R}$ . Hence the claim is true. Show that the subset of  $n \times n$  symmetric matrices is also a subspace of  $\text{Mat}_n(\mathbb{R})$ .  $\square$

**2.C.8.** Construct in Sage the 2-dimensional subspaces of  $\mathbb{Q}^3$ , generated over  $\mathbb{Q}$  by two vectors of the standard basis  $\underline{e} = \{e_1 = (1, 0, 0)^T, e_2 = (0, 1, 0)^T, e_3 = (0, 0, 1)^T\}$ .  $\circ$

*Dealing with vector spaces, it is crucial to understand the linear dependence or independence of vectors. A basis is an independent set of generators. In finite-dimensional cases, this is the same as a maximal set of linearly independent vectors. Every finite-dimensional  $\mathbb{K}$ -vector space  $V$  admits a basis; however, bases are not at*



assigning the vector  $A \cdot x$  to each vector  $x$ . If there should be a solution,  $b$  must be in the image under this mapping and thus it must be a linear combination of the columns in  $A$ .

If we extend the matrix  $A$  by the column  $b$ , the number of linearly independent columns and thus also rows might increase (but does not have to). If this number increases, then  $b$  is not in the image and the system of equations does not have a solution. If on the other hand the number of linearly independent rows does not change after adding the column  $b$  to the matrix  $A$ , it means that  $b$  must be a linear combination of the columns of  $A$ . Coefficients of such combinations are then exactly the solutions of our system.

Consider now two fixed solutions  $x$  and  $y$  of our system and some solution  $z$  of the homogeneous system with the same matrix. Then clearly

$$\begin{aligned}A \cdot (x - y) &= b - b = 0 \\ A \cdot (x + z) &= b + 0 = b.\end{aligned}$$

Thus we can summarise in the form of the so called *Kronecker-Capelli theorem*<sup>1</sup>:

#### KRONECKER-CAPELLI THEOREM

**Theorem.** *The solution of a non-homogeneous system of linear equations  $A \cdot x = b$  exists if and only if adding the column  $b$  to the matrix  $A$  does not increase the number of linearly independent rows. In such a case the space of all solutions is given by all sums of one fixed particular solution of the system and all solutions of the homogeneous system that has the same matrix.*

**2.3.6. Sums of subspaces.** Since we now have some intuition about generators and the subspaces generated by them, we should understand the possibilities of how some subspaces can generate the whole space  $V$ .



#### SUM OF SUBSPACES

Let  $V_i, i \in I$ , be subspaces of  $V$ . Then the subspace generated by their union, that is,  $\text{span} \cup_{i \in I} V_i$ , is called the *sum of subspaces*  $V_i$ . We denote it as  $W = \sum_{i \in I} V_i$ . Notably, for a finite number of subspaces  $V_1, \dots, V_k \subset V$  we write

$$W = V_1 + \cdots + V_k = \text{span}(V_1 \cup V_2 \cup \cdots \cup V_k).$$

We see that every element in the considered sum  $W$  can be expressed as a linear combination of vectors from the subspaces  $V_i$ . Because vector addition is commutative, we can

<sup>1</sup>A common formulation of this fact is “system has a solution if and only if the rank of its matrix equals the rank of its extended matrix”. Leopold Kronecker was a very influential German Mathematician, who dealt with algebraic equations in general and in particular pushed forward Number Theory in the middle of 19th century. Alfredo Capelli, an Italian, worked on algebraic identities. This theorem is equally often called by different names, e.g. *Rouché-Frobenius theorem* or *Rouché-Capelli theorem* etc. This is a very common feature in Mathematics.

all unique, cf. 2.3.7, 2.3.10. At the same time, the number of vectors in any basis is unique, and we call it the dimension  $\dim_{\mathbb{K}}(V)$  of  $V$ . We write just  $\dim(V)$  if the field  $\mathbb{K}$  is clear from the context. Notice, that vector spaces (over fixed field  $\mathbb{K}$ ) of the same finite dimension are isomorphic, cf. 2.3.13.

**2.C.9.** Consider any matrix  $A$  of size  $m \times n$  over a field  $\mathbb{K}$ . Besides the kernel  $\text{Ker}(A) = \{x \in \mathbb{K}^n : Ax = 0\}$  of  $A$ , there are several other fundamental subspaces associated with  $A$ :

- The *column space*  $C(A) = \{Ax : x \in \mathbb{K}^n\}$ , which is the span of the columns of  $A$ .
- The *row space*  $C(A^T) = \{A^T y : y \in \mathbb{K}^m\}$ , which is the span of the rows of  $A$ .
- The *left null space*  $\text{Ker}(A^T) = \{y \in \mathbb{K}^m : A^T y = 0\}$ , also known as the *cokernel* of  $A$ .

Prove that  $C(A^T)$  is a subspace of  $\mathbb{K}^n$ , while  $\text{Ker}(A^T)$  and  $C(A)$  are subspaces of  $\mathbb{K}^m$ . Moreover, show that  $\dim_{\mathbb{K}} C(A) = \text{rank}(A) = \dim_{\mathbb{K}} C(A^T)$ .  $\circ$

*Although vector spaces over infinite fields are always infinite sets, Sage can effectively handle them as mathematical objects. The main tool to this is proper work with the concept of basis. See our discussion in 2.C.14 for the finite case and see also the discussion in 2.3.10.*



**2.C.10.** Determine whether or not the vectors  $u_1 = (1, 2, 3, 1)^T$ ,  $u_2 = (1, 0, -1, 1)^T$ ,  $u_3 = (2, 1, -1, 3)^T$  and  $u_4 = (0, 0, 3, 2)^T$  are linearly independent. Do they provide a basis of  $\mathbb{R}^4$ ? Describe an answer using Sage, as well.

**Solution.** Given  $n$  vectors on  $\mathbb{R}^n$ , say

$$x_1 = (x_{11}, \dots, x_{1n})^T, \quad \dots, \quad x_n = (x_{n1}, \dots, x_{nn})^T,$$

it is easy to check that the condition  $\det(A) \neq 0$  is equivalent to their linear independence, where  $A$  is the coefficient matrix,

$$A = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{pmatrix}.$$

For the given vectors on  $\mathbb{R}^4$ , we compute

$$\det(A) = \begin{vmatrix} 1 & 2 & 3 & 1 \\ 1 & 0 & -1 & 1 \\ 2 & 1 & -1 & 3 \\ 0 & 0 & 3 & 2 \end{vmatrix} = 10 \neq 0.$$

Thus the set  $\underline{u} = (u_1, u_2, u_3, u_4)$  consists of linearly independent vectors. Since  $\dim \mathbb{R}^4 = 4$  and this set is a basis for its own linear span,  $\underline{u}$  is indeed a basis. To check the linear independence/dependence in Sage, we can rely on built-in functions, as follows:

```
u1=vector(RR, [1, 2, 3, 1])
```

aggregate summands that belong to the same subspace and for a finite sum of  $k$  subspaces we obtain

$$V_1 + V_2 + \dots + V_k = \{v_1 + \dots + v_k; v_i \in V_i, i = 1, \dots, k\}.$$

The sum  $W = V_1 + \dots + V_k \subset V$  is called the *direct sum* of subspaces if the intersection of each  $V_i$  with the sum of the other spaces is trivial, that is,  $V_i \cap \sum_{j \neq i} V_j = \{0\}$ . We show that in such a case, every vector  $w \in W$  can be written in a unique way as the sum

$$w = v_1 + \dots + v_k,$$

where  $v_i \in V_i$ . Indeed, if we could simultaneously write  $w$  as  $w = v'_1 + \dots + v'_k$ , then

$$0 = w - w = (v_1 - v'_1) + \dots + (v_k - v'_k).$$

If  $v_i - v'_i$  is the first nonzero term of the right-hand side, then this vector from  $V_i$  can be expressed using vectors from the other subspaces. This is a contradiction to the assumption that  $V_i$  has zero intersection with all the other subspaces. The only possibility is then that all the vectors on the right-hand side are zero and thus the expression of  $w$  is unique.

For direct sums of subspaces we write

$$W = V_1 \oplus \dots \oplus V_k = \bigoplus_{i=1}^k V_i.$$

**2.3.7. Basis.** Now we have everything prepared for understanding minimal sets of generators as we understood them in the plane  $\mathbb{R}^2$  and to prove the promised independence of the number of basis elements on any choices.



A basis of a  $k$ -dimensional space will usually be denoted as a  $k$ -tuple  $\underline{v} = (v_1, \dots, v_k)$  of basis vectors. This is just a matter of convention: with finitely dimensional vector spaces we shall always consider the bases along with a given order of the elements, even if we have not defined it that way (strictly speaking).

Clearly, if  $(v_1, \dots, v_n)$  is a basis of  $V$ , then the whole space  $V$  is the direct sum of the one-dimensional subspaces

$$V = \text{span}\{v_1\} \oplus \dots \oplus \text{span}\{v_n\}.$$

An immediate corollary of the derived uniqueness of decomposition of any vector  $w$  in  $V$  into the components in the direct sum gives a unique decomposition

$$w = x_1 v_1 + \dots + x_n v_n.$$

This allows us, after choosing a basis, to see the abstract vectors again as  $n$ -tuples of scalars. We shall return to this idea in paragraph 2.3.11, when we finish the discussion of the existence of bases and sums of subspaces in the general case.

**2.3.8. Theorem.** *From any finite set of generators of a vector space  $V$  we can choose a basis. Every basis of a finitely dimensional space  $V$  has the same number of elements.*

**PROOF.** The first claim is easily proved using induction on the number of generators  $k$ .

```
u2=vector(RR, [1, 0, -1, 1])
u3=vector(RR, [2, 1, -1, 3])
u4=vector(RR, [0, 0, 3, 2])
V=RR^4; V.linear_dependence([u1, u2, u3, u4])
```

Sage’s output is True. In case you want to compute only the determinant of  $A$ , then replace the last line by the syntax given here:

```
A=column_matrix([u1, u2, u3, u4])
show(det(A))
```

□

**2.C.II.** Consider the vector space  $V = \mathbb{Q}^4$  and the vectors  $v_1 = (1, 0, 1, 2)^T$  and  $v_2 = (0, 1, 0, 0)^T$ , both lying on  $V$ . What is the dimension of the vector subspace of  $V = \mathbb{Q}^4$ , spanned by  $v_1, v_2$ ? ○

**2.C.I2.** (a) Consider  $\mathbb{Q}^3$  over the rationals and let  $W$  be the subspace of  $\mathbb{Q}^3$  generated by the vectors  $v_1 = (2, -1, 3)^T$  and  $v_2 = (4, 0, 1)^T$ . Determine the expression of a general vector of  $W$  and next use Sage to determine the dimension of  $W$ . What happens if we replace  $v_2$  by a scalar multiple of  $v_1$ ? (b) Consider  $\mathbb{Q}^4$  over the rationals and let  $W$  be the subspace spanned by the vectors  $u_1 = (1, 1, 0, 0)^T$  and  $u_2 = (0, 1, 1, 0)^T$ . Describe the *quotient space*  $V/W$  (i.e., “what is left” in  $V$  viewing vectors “up to  $W$ ”, cf. 3.4.13) and next verify your answer via Sage. ○

**2.C.I3.** For the second task in 2.C.12, find a basis of the quotient space  $V/W$ . Confirm your answer by Sage by using the method `vector` in `subspace`. ○

**2.C.I4.** (a) Show that the following set of vectors is a subspace of  $\mathbb{Z}_5^3$ ,

$$U = \left\{ \begin{pmatrix} a \\ 2b \\ 3a + 4b \end{pmatrix} : a, b \in \mathbb{Z}_5 \right\}.$$

(b) Find the reduced row echelon form of the matrix  $A$  having as columns the generators of  $U$ . Next deduce that  $\dim_{\mathbb{Z}_5} U = 2$  by finding a basis of  $U$ . ○

*We can often rely on Sage to perform mathematical “experiments”. In linear algebra for example, this can be done using matrices or vectors with random coefficients. Notice that such experiments can be useful to test theoretical statements, or illustrate them in a random way. Let us describe such an example.*

**2.C.15.** Provide in Sage a short code that generates a random (ordered) tuple  $S$  of vectors within  $V = \mathbb{Q}^4$ , and checks their linear independence. Moreover, create a random  $4 \times 4$  matrix  $A$  whose columns are the vectors in  $S$ , and use the matrix’s rank to establish an alternative criterium for the linear independence of  $S$  (recall the following basic criterium for the invertibility of a square matrix  $A \in \text{Mat}_n(\mathbb{K})$ :  $A$  is invertible if and only if  $\text{rank}(A) = n$ ).

Only the zero subspace does not need a generator and thus we are able to choose an empty basis. On the other hand, we are not able to choose the zero vector (the generators would then be linearly dependent) and there is nothing else in the subspace.

In order to have our inductive step more natural, we deal with the case  $k = 1$  first. We have  $V = \text{span}\{v\}$  and  $v \neq 0$ , because  $\{v\}$  is a linearly independent set of vectors. Then  $\{v\}$  is also a basis of the vector space  $V$  and any other vector is a multiple of  $v$ , so all bases of  $V$  must contain exactly one vector, which can be chosen from any set of generators.

Assume that the claim holds for  $k = n$  and consider  $V = \text{span}\{v_1, \dots, v_{n+1}\}$ . If  $v_1, \dots, v_{n+1}$  are linearly independent, then they form a basis. If they are linearly dependent, there exists  $i$  such that

$$v_i = a_1 v_1 + \dots + a_{i-1} v_{i-1} + a_{i+1} v_{i+1} + \dots + a_{n+1} v_{n+1}.$$

Then  $V = \text{span}\{v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_{n+1}\}$  and we can choose a basis, using the inductive assumption.

It remains to show that bases always have the same number of elements. Consider a basis  $\underline{v} = (v_1, \dots, v_n)$  of the space  $V$  and for an arbitrary nonzero vector  $u$ , consider

$$u = a_1 v_1 + \dots + a_n v_n \in V$$

with  $a_i \neq 0$  for some  $i$ . Then

$$v_i = \frac{1}{a_i} (u - (a_1 v_1 + \dots + a_{i-1} v_{i-1} + a_{i+1} v_{i+1} + \dots + a_n v_n))$$

and therefore also  $\text{span}\{u, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n\} = V$ .

We show that this is again a basis. For if adding  $u$  to the linearly independent vectors  $v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n$  leads to a set of linearly dependent vectors, then

$$V = \text{span}\{v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n\},$$

which implies a basis of  $n - 1$  vectors chosen from  $\underline{v}$ , which is not possible.

Thus we have proved that for any nonzero vector  $u \in V$  there exists  $i$ ,  $1 \leq i \leq n$ , such that  $(u, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n)$  is again a basis of  $V$ .

Similarly, instead of one vector  $u$ , we can consider a linearly independent set  $u_1, \dots, u_k$ . We will sequentially add  $u_1, u_2, \dots$ , always exchanging for some  $v_i$  using our previous approach. We have to ensure that there always is such  $v_i$  to be replaced (that is, that the vectors  $u_i$  will not consequently replace each other).

Assume thus that we have already placed  $u_1, \dots, u_\ell$  instead of some  $v_j$ ’s. Then the vector  $u_{\ell+1}$  can be expressed as a linear combination of the latter vectors  $u_i$  and the remaining  $v_j$ ’s. As we have seen,  $u_{\ell+1}$  may replace any vector with non-zero coefficient in this linear combination. If only the coefficients at  $u_1, \dots, u_\ell$  were nonzero, then it would mean that the vectors  $u_1, \dots, u_{\ell+1}$  were linearly dependent, which is a contradiction.

Summarizing, for every  $k \leq n$  we can arrive after  $k$  steps at a basis in which  $k$  vectors from the original basis were exchanged for the new  $u_i$ ’s. If  $k > n$ , then in the  $n$ -th step we

**Solution.** To introduce a random vector in Sage we can use the command `V.random_element()`, where  $V$  is the underlying vector space. For instance, the cell

```
V=QQ^4
u=V.random_element(); show(u)
```

prints a random vector of  $\mathbb{Q}^4$ . Any time that the code will be implemented, a new vector will appear in our screen. To construct simultaneously four random vectors lying on  $\mathbb{Q}^4$  and check their linear independence, we can instead type

```
V=QQ^4
S=[V.random_element() for i in range(4)]
show(S); V.linear_dependence(S)==[]
```

which will return an ordered random tuple  $S$  of vectors in  $\mathbb{Q}^4$  and moreover the message “True”, or “False”, depending on the linear independence/dependence of  $S$ . Or type

```
V=QQ^4
S=[V.random_element() for i in range(4)]
W=V.span(S); print(W)
```

In this case, if Sage confirms that  $W$  is a 4-dimensional subspace, then the set  $S$  of four random vectors should be linearly independent. Remarkably, most of the cases Sage returns a linearly independent set, and hence this method can be used when you need to locate many different (ordered) bases of a given vector space. We recommend practicing with further examples in your editor.

Finally, to introduce the required matrix  $A$  and specify a rank-based criterium for linear independence or dependence, you can add the following code in either of the previous two cells:

```
A=column_matrix(S)
if rank(A) < 4 :
    # then linear dependent
        print(False)
else :
    # then linear independent
        print(True)
```

Thus, when  $A$  is of full rank, i.e.,  $\text{rank}(A) = 4$ ,  $S$  will consist of linearly independent vectors, and Sage will type `True`. When  $\text{rank}(A) < 4$ ,  $S$  will consist of linearly dependent vectors, and its message will be `False`.  $\square$

**2.C.16.** Given arbitrary linearly independent vectors  $u, v, w, z$  in a vector space  $V$ , decide whether or not the vectors

$$u - 2v, \quad 3u + w - z, \quad u - 4v + w + 2z, \quad 4v + 8w + 4z$$

are also linearly independent.  $\circ$

**2.C.17.** Determine the coordinates of the vector  $w = (1, 1, 1)^T$  with respect to the following ordered basis of  $\mathbb{R}^3$ :

$$\underline{u} = \{u_1 = (1, 2, 1)^T, u_2 = (-1, 1, 0)^T, u_3 = (0, 1, 1)^T\}.$$

Next, verify your answer in Sage, based for example on the command `coordinates`, or any other method.

would obtain a basis consisting only of new vectors  $u_i$ , which means that the original set could not be linearly independent.

In particular, it is not possible for two bases to have a different number of elements.  $\square$

In fact, we have proved a much stronger claim, the *Steinitz exchange lemma*:

STEINITZ EXCHANGE LEMMA

For every finite basis  $\underline{v}$  of a vector space  $V$  and every set of linearly independent vectors  $u_i, i = 1, \dots, k$  in  $V$  we can find a subset of the basis vectors  $v_j$  which will complete the set of  $u_i$ 's into a new basis.

**2.3.9. Corollaries of the Steinitz lemma.** Because of the



possibility of freely choosing and replacing basis vectors we can immediately derive nice (and intuitively expectable) properties of bases of vector spaces:

**Proposition.** (1) Every two bases of a finite dimensional vector space have the same number of elements, that is, our definition of dimension is basis-independent.

(2) If  $V$  has a finite basis, then every linearly independent set can be extended to a basis.

(3) A basis of a finite dimensional vector space is a maximal linearly independent set of vectors.

(4) The bases of a vector space are the minimal sets of generators.

A little more complicated, but now easy to deal with, is the situation of dimensions of subspaces and their sums:

**Corollary.** Let  $W, W_1, W_2 \subset V$  be subspaces of a space  $V$  of finite dimension. Then

- (1)  $\dim W \leq \dim V$ ,
- (2)  $V = W$  if and only if  $\dim V = \dim W$ ,
- (3)  $\dim W_1 + \dim W_2 = \dim(W_1 + W_2) + \dim(W_1 \cap W_2)$ .



**PROOF.** It remains to prove only the last claim. This is evident if the dimension of one of the spaces is zero. Assume  $\dim W_1 = r \geq 1$ ,  $\dim W_2 = s \geq 1$  and let  $(w_1 \dots, w_t)$  be a basis of  $W_1 \cap W_2$  (or empty set, if the intersection is trivial).

According to the Steinitz exchange lemma this basis of the intersection can be extended to a basis  $(w_1, \dots, w_t, u_{t+1}, \dots, u_r)$  for  $W_1$  and to a basis  $(w_1 \dots, w_t, v_{t+1}, \dots, v_s)$  for  $W_2$ . Vectors

$$w_1, \dots, w_t, u_{t+1}, \dots, u_r, v_{t+1}, \dots, v_s$$

clearly generate  $W_1 + W_2$ . We show that they are linearly independent. Let

$$a_1 w_1 + \dots + a_t w_t + b_{t+1} u_{t+1} + \dots + \dots + b_r u_r + c_{t+1} v_{t+1} + \dots + c_s v_s = 0.$$



**Solution.** When a vector  $u \in \mathbb{R}^n$  is given in coordinates, such as  $w$ , it is always understood to be expressed with respect to the standard basis, unless stated otherwise. The assumption that  $\underline{u}$  is an ordered basis, with a fixed and well defined order as indicated by the relevant positions of the vectors  $u_1, u_2, u_3$ , it is also crucial. To solve the problem we need to determine reals  $a, b$ , and  $c \in \mathbb{R}$ , which satisfy the following matrix equation:

$$a \cdot \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + b \cdot \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

This is equivalent to  $(a - b, 2a + b + c, a + c)^T = (1, 1, 1)^T$ , which induces the following system of equations:

$$\{a - b = 1, \quad 2a + b + c = 1, \quad a + c = 1\}.$$

It is easy to see that this system has a unique solution, given by  $a = \frac{1}{2}, b = -\frac{1}{2}, c = \frac{1}{2}$ . Thus, the coordinates of  $w$  with respect to  $\underline{u}$  are given by  $(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2})$ .

There are several different methods to demonstrate a verification via Sage. Here we describe two methods. The first method uses the `coordinates` command, which is designed to determine the coordinates of a vector with respect to a given basis. Its implementation goes as follows:

```
V=RR^3
u1=vector([1, 2, 1]); u2=vector([-1, 1, 0])
u3=vector([0, 1, 1]); w=vector([1, 1, 1])
L=[u1, u2, u3]
W=V.subspace_with_basis(L)
cord=W.coordinates(w); show(cord)
sum([cord[i]*L[i] for i in range(3)])==w
```

This cell returns the coefficients of  $w$  with respect to  $\underline{u}$ , along with the message `True`. The latter indicates that the command in the last row verifies the accuracy of the result from `show(cord)` (i.e., Sage is used to verify its own output).

The second method relies on computing the reduced row echelon form of the matrix having as columns the vectors of the given basis, together with  $w$ , and so of the extended matrix corresponding to the linear system posed above. For this, add in the previous cell the code

```
B=column_matrix([u1, u2, u3, w])
B1=B.rref(); show(B1[:, 3])
```

By executing the block, Sage prints out the fourth column of the reduced row echelon form, which indeed consists of the coefficients of  $w$  with respect to the ordered basis  $\underline{u}$ .  $\square$



*So far we have demonstrated a straightforward algorithm to identify a maximal linearly independent set from a given collection of vectors. The steps are as follows:*

- Arrange the given vectors as columns in a matrix.
- Transform the matrix into reduced row echelon form using row operations.
- Identify the vectors corresponding to the pivot columns; these form a maximal linearly independent set.

Then necessarily

$$\begin{aligned} &-(c_{t+1} \cdot v_{t+1} + \dots + c_s \cdot v_s) = \\ &= a_1 \cdot w_1 + \dots + a_t \cdot w_t + b_{t+1} \cdot u_{t+1} + \dots + b_r \cdot u_r \end{aligned}$$

must belong to  $W_2 \cap W_1$ . This implies that

$$b_{t+1} = \dots = b_r = 0,$$

since this is the way we have defined our bases. Then also

$$a_1 \cdot w_1 + \dots + a_t \cdot w_t + c_{t+1} \cdot v_{t+1} + \dots + c_s \cdot v_s = 0$$

and because the corresponding vectors form a basis  $W_2$ , all the coefficients are zero.

The claim (3) now follows by directly counting the generators.  $\square$

**2.3.10. Examples.** (1)  $\mathbb{K}^n$  has (as a vector space over  $\mathbb{K}$ ) dimension  $n$ . The  $n$ -tuple of vectors

$$((1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, \dots, 0, 1))$$

is clearly a basis, called the *standard basis* of  $\mathbb{K}^n$ .

Note that in the case of a finite field of scalars, say  $\mathbb{Z}_k$  with  $k$  prime, the whole space  $\mathbb{K}^n$  has only a finite number  $k^n$  of elements.

(2)  $\mathbb{C}$  as a vector space over  $\mathbb{R}$  has dimension 2. A basis is for instance the pair of numbers 1 and  $i$ , or any other two complex numbers which are not a real multiple of each other, eg.  $1 + i$  and  $1 - i$ .

(3)  $\mathbb{K}_m[x]$ , that is, the space of all polynomials with coefficients in  $\mathbb{K}$  of degree at most  $m$ , has dimension  $m + 1$ . A basis is for instance the sequence  $1, x, x^2, \dots, x^m$ .

The vector space of all polynomials  $\mathbb{K}[x]$  has dimension  $\infty$ , but we can still find a basis (although infinite in size):  $1, x, x^2, \dots$ .

(4) The vector space  $\mathbb{R}$  over  $\mathbb{Q}$  has dimension  $\infty$ . It does not have a countable basis.

(5) The vector space of all mappings  $f : \mathbb{R} \rightarrow \mathbb{R}$  has also dimension  $\infty$ . It does not have any countable basis.

**2.3.11. Vector coordinates.** If we fix a basis  $(v_1, \dots, v_n)$  of a finite dimensional space  $V$ , then every vector  $w \in V$  can be expressed as a linear combination  $w = a_1 v_1 + \dots + a_n v_n$  in a unique way. Indeed, assume that we can do it in two ways:



$$w = a_1 v_1 + \dots + a_n v_n = b_1 v_1 + \dots + b_n v_n.$$

Then

$$0 = (a_1 - b_1) \cdot v_1 + \dots + (a_n - b_n) \cdot v_n$$

and thus  $a_i = b_i$  for all  $i = 1, \dots, n$ , because the vectors  $v_i$  are linearly independent. We have reached the concept of coordinates:

This approach is essentially what was utilized in the second method described in 2.C.17. For a theoretical validation and a more in-depth discussion, see the theoretical section 2.3.5.

**2.C.18.** Determine the vector subspace of the Euclidean space  $\mathbb{R}^4$ , generated by the vectors  $u_1 = (-1, 3, -2, 1)^T$ ,  $u_2 = (2, -1, -1, 2)^T$ ,  $u_3 = (-4, 7, -3, 0)^T$ , and  $u_4 = (1, 5, -5, 4)^T$ .

**Solution.** Write the vectors  $u_i$  into the columns of a matrix and transform it using elementary row transformations. This gives

$$\begin{pmatrix} -1 & 2 & -4 & 1 \\ 3 & -1 & 7 & 5 \\ -2 & -1 & -3 & -5 \\ 1 & 2 & 0 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ -1 & 2 & -4 & 1 \\ 3 & -1 & 7 & 5 \\ -2 & -1 & -3 & -5 \end{pmatrix} \\ \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 4 & -4 & 5 \\ 0 & -7 & 7 & -7 \\ 0 & 3 & -3 & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 1 & -1 & 5/4 \\ 0 & 0 & 0 & -1/4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ \sim \begin{pmatrix} 1 & 2 & 0 & 4 \\ 0 & 1 & -1 & 5/4 \\ 0 & 0 & 0 & -1/4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Now, according to the algorithm above, a maximal linearly independent set (basis) consists of those vectors corresponding to the columns with the pivots. For our case, the pivots are circled in the reduced row echelon form<sup>8</sup> and hence a maximal linearly independent set consists of the vectors  $u_1$ ,  $u_2$  and  $u_4$ . This means that the vector subspace generated by  $\{u_1, u_2, u_3, u_4\}$  is only 3-dimensional.

Finally, let us verify the output of our algorithm via Sage (where for convenience we view the vectors as elements of  $\mathbb{Q}^4$ ). Let us check first that the given vector are indeed linearly dependent:

```
v1=vector(QQ, [-1, 3, -2, 1])
v2=vector(QQ, [2, -1, -1, 2])
v3=vector(QQ, [-4, 7, -3, 0])
v4=vector(QQ, [1, 5, -5, 4])
V=QQ^4; L=[v1, v2, v3, v4]
V.linear_dependence(L)==[]
```

Next we apply the trick with the pivot elements:

```
M=column_matrix([v1, v2, v3, v4])
M1=M.rref(); show(M1); M.pivots()
```

with the last command returning the desired  $(0, 1, 3)$ . □

**2.C.19.** Derive in Sage a method analogous to the one presented in 2.C.18, with aim to complete the following set of vectors to a basis of the vector space  $\mathbb{Q}^4$ :

$$L = \{v_1 = (2, 0, 1, 0)^T, v_2 = (1, 0, 1, 0)^T\}.$$

<sup>8</sup>We will adopt this tactic especially in Chapter 3.

COORDINATES OF VECTORS

**Definition.** The coefficients of the unique linear combination expressing the given vector  $w \in V$  in the chosen basis  $\underline{v} = (v_1, \dots, v_n)$  are called the *coordinates of the vector  $w$*  in this basis.

Whenever we speak about coordinates  $(a_1, \dots, a_n)$  of a vector  $w$ , which we express as a sequence, we must have a fixed ordering of the basis vectors  $\underline{v} = (v_1, \dots, v_n)$ . Although we have defined the basis as a minimal set of generators, in reality we work with them as with sequences (that is, with completely ordered sets).

ASSIGNING COORDINATES TO VECTORS

A mapping assigning the vector  $v = a_1v_1 + \dots + a_nv_n$  to its coordinates in the basis  $\underline{v}$  will be denoted by the same symbol  $\underline{v} : V \rightarrow \mathbb{K}^n$ . It has the following properties:

- (1)  $\underline{v}(u + w) = \underline{v}(u) + \underline{v}(w); \forall u, w \in V,$
- (2)  $\underline{v}(a \cdot u) = a \cdot \underline{v}(u); \forall a \in \mathbb{K}, \forall u \in V.$

Note that the operations on the two sides of these equations are not identical. Quite the opposite; they are operations on different vector spaces!

Sometimes it is really useful to understand vectors as mappings from fixed set of independent generators to coordinates (without having the generators ordered). In this way, we may think about the basis  $M$  of infinite dimensional vector spaces  $V$ . Even though the set  $M$  will be infinite, there can be only a finite number of non-zero values for any mapping representing a vector. The vector space of all polynomials  $\mathbb{K}_\infty[x]$ , with the basis  $M = \{1, x, x^2, \dots\}$  is a good example.

**2.3.12. Linear mappings.** The above properties of the assignments of coordinates are typical for what we have called linear mappings in the geometry of the plane  $\mathbb{R}^2$ .

For any vector space (of finite or infinite dimension) we define “linearity” of a mapping between spaces in a similar way to the case of the plane  $\mathbb{R}^2$ :

LINEAR MAPPINGS

Let  $V$  and  $W$  be vector spaces over the same field of scalars  $\mathbb{K}$ . The mapping  $f : V \rightarrow W$  is called a *linear mapping*, or *homomorphism*, if the following holds:

- (1)  $f(u + v) = f(u) + f(v), \forall u, v \in V$
- (2)  $f(a \cdot u) = a \cdot f(u), \forall a \in \mathbb{K}, \forall u \in V.$

We have seen such mappings already in the case of matrix multiplication:

$$f : \mathbb{K}^n \rightarrow \mathbb{K}^m, \quad x \mapsto A \cdot x$$

with a fixed matrix  $A$  of the type  $m/n$  over  $\mathbb{K}$ .

Analogously to the abstract definition of vector spaces, it is again necessary to prove seemingly trivial claims that follow from the axioms:

**2.C.20.** Consider the matrix  $A = \begin{pmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 5 \end{pmatrix}$ .

- i) Find the column space  $C(A)$ , the row space  $C(A^T)$ , the kernel  $\text{Ker}(A)$  and the cokernel  $\text{Ker}(A^T)$  of  $A$ .
- ii) Compute the dimensions of these subspaces, using Sage as well. Hint: Use the commands `column_space`, `row_space`, `right_kernel` and `left_kernel`.

**Solution.** (i) The given matrix  $A$  has size  $3 \times 2$  and thus the column space should be a subspace of  $\mathbb{R}^3$ . Let us check the linear independency of the two column vectors of  $A$ . By performing elementary row operations we see that

$$A = \begin{pmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 5 \end{pmatrix} \xrightarrow[R_3 \rightarrow \frac{1}{5}R_3]{R_1 \rightarrow \frac{1}{2}R_1} \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - R_1} \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \xrightarrow{R_3 \rightarrow R_3 - R_2} \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \xrightarrow{R_1 \rightarrow R_1 - 2R_2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Thus  $\text{rank}(A) = 2$  and the column vectors are linearly independent. Hence,  $C(A) = \text{span}\{(2, 1, 0)^T, (4, 3, 5)^T\}$  with  $\dim_{\mathbb{R}} C(A) = 2$ .

By the reduced row echelon form of  $A$  we also deduce that the first two rows of  $A$  are linearly independent. For instance, it is easy to see that the third row is a linear combination of the previous two:  $(0, 5)^T = a(2, 4)^T + b(1, 3)^T$  with  $a = -5/2$  and  $b = 5$ . Thus,  $C(A^T) = \text{span}\{(2, 4)^T, (1, 3)^T\}$  with  $\dim_{\mathbb{R}} C(A^T) = 2$ , as well.

Recall that null spaces are the solution spaces of linear homogeneous systems  $Au = 0$ . For our case, this system has the form

$$\begin{pmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and it is easy to see that the unique solution is given by  $x_1 = 0 = x_2$ . Thus,  $\text{Ker}(A)$  is trivial,  $\text{Ker}(A) = \{(0, 0)^T\} \subset \mathbb{R}^2$ . In other words the linear map  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^3$  corresponding to  $A$  is an injection.

On the other hand, the left null space is the kernel of the transpose matrix  $A^T$ . This space consists of vectors  $w = (w_1, w_2, w_3)^T \in \mathbb{R}^3$ , such that  $A^T w = 0$ , i.e.,

$$\begin{pmatrix} 2 & 1 & 0 \\ 4 & 3 & 5 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Solving the corresponding system we obtain  $\text{Ker}(A^T) = \text{span}\{(5, -10, 2)^T\} \subset \mathbb{R}^3$  and hence the cokernel of  $A$  is 1-dimensional.

(ii) In Sage, you can easily execute the commands outlined in the statement as follows:

```
A=matrix([[2, 4], [1, 3], [0, 5]])
rank(A)
print(A.column_space())
print(dim(A.column_space())==rank(A))
print(A.row_space())
```

**Proposition.** Let  $f : V \rightarrow W$  be a linear mapping between two vector spaces over the same field of scalars  $\mathbb{K}$ . The following is true for all vectors  $u, u_1, \dots, u_k \in V$  and scalars  $a_1, \dots, a_k \in \mathbb{K}$

- (1)  $f(0) = 0$ ,
- (2)  $f(-u) = -f(u)$ ,
- (3)  $f(a_1 \cdot u_1 + \dots + a_k \cdot u_k) = a_1 \cdot f(u_1) + \dots + a_k \cdot f(u_k)$ ,
- (4) for every vector subspace  $V_1 \subset V$ , its image  $f(V_1)$  is a vector subspace in  $W$ ,
- (5) for every vector subspace  $W_1 \subset W$ , the set  $f^{-1}(W_1) = \{v \in V; f(v) \in W_1\}$  is a vector subspace in  $V$ .

**PROOF.** We rely on the axioms, definitions and already proved results (in case you are not sure what has been used, look it up!):

$$f(0) = f(u - u) = f((1 - 1) \cdot u) = 0 \cdot f(u) = 0,$$

$$f(-u) = f((-1) \cdot u) = (-1) \cdot f(u) = -f(u).$$

Property (3) is derived easily from the definition for two summands, using induction on the number of summands.

Next, (3) implies  $\text{span } f(V_1) = f(V_1)$ , thus it is a vector subspace. On the other hand, if  $f(u) \in W_1$  and  $f(v) \in W_1$  then for any scalars we arrive at  $f(a \cdot u + b \cdot v) = a \cdot f(u) + b \cdot f(v) \in W_1$ .  $\square$

The *image* of a linear mapping,  $\text{Im } f = f(V) \subset W$ , is always a vector subspace, since for any set of vectors  $u_i$ , the linear combination of images  $f(u_i)$  is the image of the linear combination of the vectors  $u_i$  with the same coefficients.

Analogously, the set of all vectors  $\text{Ker } f = f^{-1}(\{0\}) \subset V$  is a subspace, since the linear combination of zero images will always be a zero vector. The subspace  $\text{Ker } f$  is called the *kernel of the linear mapping*  $f$ .

A linear mapping which is a bijection is called an *isomorphism*.

**2.3.13. Proposition** (Simple corollaries). (1) The composition  $g \circ f : V \rightarrow Z$  of two linear mappings  $f : V \rightarrow W$  and  $g : W \rightarrow Z$  is again a linear mapping.

(2) The linear mapping  $f : V \rightarrow W$  is an isomorphism if and only if  $\text{Im } f = W$  and  $\text{Ker } f = \{0\} \subset V$ . The inverse mapping of an isomorphism is again an isomorphism.

(3) For any two subspaces  $V_1, V_2 \subset V$  and linear mapping  $f : V \rightarrow W$ ,

$$f(V_1 + V_2) = f(V_1) + f(V_2),$$

$$f(V_1 \cap V_2) \subset f(V_1) \cap f(V_2).$$

(4) The “coordinate assignment” mapping  $\underline{u} : V \rightarrow \mathbb{K}^n$  given by an arbitrarily chosen basis  $\underline{u} = (u_1, \dots, u_n)$  of a vector space  $V$  is an isomorphism.

(5) Two finitely dimensional vector spaces are isomorphic if and only if they have the same dimension.

(6) The composition of two isomorphisms is an isomorphism.

```
print(dim(A.row_space())==rank(A))
print(A.right_kernel())
print((A.T).right_kernel())
print(A.left_kernel())
print((A.T).right_kernel()==A.left_kernel())
```

In this block the final line confirms that the command `(A.T).right_kernel()` can be used as an alternative to `A.left_kernel()`.  $\square$

*In linear algebra, there are several classical constructions that generate new vector spaces from existing ones. Examples include intersections, direct sums (see e.g., 2.3.6), and quotients of vector spaces. While intersection means just solving systems of linear equations on coordinates, sums require merging all generators (and leaving only a maximal independent subset). We already touched quotient spaces in 2.C.12, although they are formally introduced in Chapter 3, see 3.4.13. Since we are short in space, we display only a few tasks and further discussion appears later, see also Section E.*



**2.C.21.** In  $\mathbb{R}^4$  consider the three-dimensional linear subspaces  $U = \text{span}\{u_1, u_2, u_3\}$  and  $V = \text{span}\{v_1, v_2, v_3\}$ , where the vectors  $u_i$  and  $v_i$  ( $i = 1, 2, 3$ ) are respectively given by

$$\begin{aligned} u_1 &= (1, 1, 1, 0)^T, & v_1 &= (1, 1, -1, -1)^T, \\ u_2 &= (1, 1, 0, 1)^T, & v_2 &= (1, -1, 1, -1)^T, \\ u_3 &= (1, 0, 1, 1)^T, & v_3 &= (1, -1, -1, 1)^T. \end{aligned}$$

Determine explicitly the subspace  $W := U \cap V$ , and verify your answer via Sage. Moreover, compute  $\dim(W)$  over  $\mathbb{R}$ .

**Solution.** The intersection  $U \cap V$  contains exactly these vectors which are linear combinations of the vectors  $u_i$ , and also linear combinations of the vectors  $v_i$ , for  $i = 1, 2, 3$ . Thus, we search for some  $x_1, x_2, x_3, y_1, y_2, y_3 \in \mathbb{R}$ , satisfying

$$x_1u_1 + x_2u_2 + x_3u_3 = y_1v_1 + y_2v_2 + y_3v_3.$$

This means that we are looking for a solution of the following system of linear equations:

$$\begin{cases} x_1 + x_2 + x_3 = y_1 + y_2 + y_3, \\ x_1 + x_2 = y_1 - y_2 - y_3, \\ x_1 + x_3 = -y_1 + y_2 - y_3, \\ x_2 + x_3 = -y_1 - y_2 + y_3. \end{cases}$$

We convert this to a homogeneous system of linear equations, by moving all variables to the left-hand side. Its matrix reads as

$$A = \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 0 & -1 & 1 & 1 \\ 1 & 0 & 1 & 1 & -1 & 1 \\ 0 & 1 & 1 & 1 & 1 & -1 \end{pmatrix}.$$

Let us now apply row operations to find an echelon form:

$$A \sim \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 0 & -1 & 0 & 2 & 2 \\ 0 & -1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 1 & 1 & 1 & -1 \end{pmatrix}$$

**PROOF.** Proving the first claim is a very easy exercise left to the reader. In order to verify (2), notice that  $f$  is surjective if and only if  $\text{Im } f = W$ . If  $\text{Ker } f = \{0\}$  then  $f(u) = f(v)$  ensures  $f(u - v) = 0$ , that is,  $u = v$ . In this case  $f$  is injective. Finally, if  $f$  is a linear bijection, then the vector  $w$  is the preimage of a linear combination  $au + bv$ , that is  $w = f^{-1}(au + bv)$ , if and only if



$$f(w) = au + bv = f(a \cdot f^{-1}(u) + b \cdot f^{-1}(v)).$$

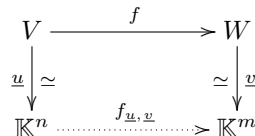
Thus we also get  $w = af^{-1}(u) + bf^{-1}(v)$  and therefore the inversion of a linear bijection is again a linear bijection.

The third property is obvious from the definition, but try finding an example showing that the inequality in the second equation can indeed be sharp.

The remaining claims all follow immediately from the definition.  $\square$

**2.3.14. Coordinates again.** Consider any two vector spaces

$V$  and  $W$  over  $\mathbb{K}$  with  $\dim V = n$ ,  $\dim W = m$  and consider some linear mapping  $f : V \rightarrow W$ . For every choice of basis  $\underline{u} = (u_1, \dots, u_n)$  on  $V$ ,  $\underline{v} = (v_1, \dots, v_m)$  on  $W$  there are the following linear mappings as shown in the diagram:



The bottom arrow  $f_{\underline{u}, \underline{v}}$  is defined by the remaining three, i.e. the composition of linear mappings

$$f_{\underline{u}, \underline{v}} = \underline{v} \circ f \circ \underline{u}^{-1}.$$

#### MATRIX OF A LINEAR MAPPING

Every linear mapping is uniquely determined by its values on an arbitrary set of generators, in particular, on the vectors of a basis  $\underline{u}$ . Denote by

$$f(u_1) = a_{11} \cdot v_1 + a_{21} \cdot v_2 + \dots + a_{m1}v_m$$

$$f(u_2) = a_{12} \cdot v_1 + a_{22} \cdot v_2 + \dots + a_{m2}v_m$$

$\vdots$

$$f(u_n) = a_{1n} \cdot v_1 + a_{2n} \cdot v_2 + \dots + a_{mn}v_m,$$

that is, scalars  $a_{ij}$  form a matrix  $A$ , where the columns are coordinates of the values  $f(u_j)$  of the mapping  $f$  on the basis vectors expressed in the basis  $\underline{v}$  on the target space  $W$ .

A matrix  $A = (a_{ij})$  is called the *matrix of the mapping*  $f$  in the bases  $\underline{u}, \underline{v}$ .

$$\begin{aligned} &\sim \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 1 & 1 & 1 & 1 & -1 \\ 0 & 0 & -1 & 0 & 2 & 2 \\ 0 & 0 & 1 & 3 & 1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 0 & 1 & 1 & 1 & 1 & -1 \\ 0 & 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \\ &\sim \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & -2 & -2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}. \end{aligned}$$

The final matrix is in row echelon form and gives that

$$x_1 = -2t, x_2 = -2s, x_3 = 2s + 2t, y_1 = -s - t, y_2 = s, y_3 = t, \quad t, s \in \mathbb{R}.$$

One can verify the previous computation in Sage, by typing

```
A=matrix(QQ, [[1, 1, 1, -1, -1, -1],
              [1, 1, 0, -1, 1, 1],
              [1, 0, 1, 1, -1, 1],
              [0, 1, 1, 1, 1, -1]])
show(A.echelon_form())
```

Hence we can claim that a general vector of  $W$  is written as

$$\begin{pmatrix} x_1 + x_2 + x_3 \\ x_1 + x_2 \\ x_1 + x_3 \\ x_2 + x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -2t - 2s \\ 2s \\ 2t \end{pmatrix},$$

that is,  $W = \text{span}_{\mathbb{R}}\{(0, -2t-2s, 2s, 2t)^T \mid s, t \in \mathbb{R}\}$ . Obviously,  $W$  is generated by the vectors  $w_1 = (0, -1, 1, 0)^T$  and  $w_2 = (0, -1, 0, 1)^T$  and it is easy to see that they are linearly independent. Thus  $\dim_{\mathbb{R}}(W) = 2$

In Sage we can determine the intersection  $U \cap V$  of two linear subspaces of a given vector space via the command `U.intersection(V)`. For our example, give the cell

```
u1=vector([1, 1, 1, 0])
u2=vector([1, 1, 0, 1])
u3=vector([1, 0, 1, 1])
U=(RR**4).span([u1, u2, u3], QQ)
v1=vector([1, 1, -1, -1])
v2=vector([1, -1, 1, -1])
v3=vector([1, -1, -1, 1])
V=(RR**4).span([v1, v2, v3], QQ)
W=U.intersection(V)
show(W); print(dim(W))
```

**2.C.22.** Referring to the results presented in 2.C.20, demonstrate that for the given matrix  $A$ , the following direct sum decompositions are valid:

$$\mathbb{R}^3 = C(A) \oplus \text{Ker}(A^T), \quad \mathbb{R}^2 = C(A^T) \oplus \text{Ker}(A).$$

**Solution.** In 2.C.20 we obtained the expressions

$$C(A) = \text{span}\{u_1 = (2, 1, 0)^T, u_2 = (4, 3, 5)^T\},$$

$$\text{Ker}(A^T) = \text{span}\{u_3 = (5, -10, 2)^T\}.$$

In particular, we have seen that  $\dim C(A) + \dim \text{Ker}(A^T) = 2 + 1 = 3 = \dim \mathbb{R}^3$ . This, along with the relation

$$C(A) \cap \text{Ker}(A^T) = \{0\}$$

that we are going to prove below, precisely defines the conditions for a direct sum, i.e.,  $\mathbb{R}^3 = C(A) \oplus \text{Ker}(A^T)$ . Hence, suppose that  $v \in C(A) \cap \text{Ker}(A^T)$ . Then, there exists scalars  $a, b, c \in \mathbb{R}$  such that  $v = au_1 + bu_2$  and  $v = cu_3$ , respectively,

For a general vector  $u = x_1u_1 + \dots + x_nu_n \in V$  we calculate (recall that vector addition is commutative and distributive with respect to scalar multiplication)

$$\begin{aligned} f(u) &= x_1f(u_1) + \dots + x_nf(u_n) \\ &= x_1(a_{11}v_1 + \dots + a_{m1}v_m) + \dots + x_n(a_{1n}v_1 + \dots) \\ &= (x_1a_{11} + \dots + x_na_{1n})v_1 + \dots + (x_1a_{m1} + \dots)v_m. \end{aligned}$$

Using matrix multiplication we can now very easily and clearly write down the values of the mapping  $f_{\underline{u}, \underline{v}}(w)$  defined uniquely by the previous diagram. Recall that vectors in  $\mathbb{K}^\ell$  are understood as columns, that is, matrices of the type  $\ell/1$

$$f_{\underline{u}, \underline{v}}(\underline{u}(w)) = \underline{v}(f(w)) = A \cdot \underline{u}(w).$$

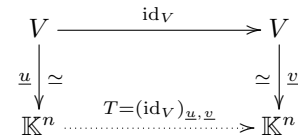
On the other hand, if we have fixed bases on  $V$  and  $W$ , then every choice of a matrix  $A$  of the type  $m/n$  gives a unique linear mapping  $\mathbb{K}^n \rightarrow \mathbb{K}^m$  and thus also a mapping  $f : V \rightarrow W$ . We have found the bijective correspondence between matrices of the fixed types (determined by dimensions of  $V$  and  $W$ ) and linear mappings  $V \rightarrow W$ .

**2.3.15. Coordinate transition matrix.** If we choose  $V =$



$W$  to be the same space, but with two different bases  $\underline{u}, \underline{v}$ , and consider the identity mapping for  $f$ , then the approach from the previous paragraph expresses the vectors of the basis  $\underline{u}$  in coordinates with respect to the basis  $\underline{v}$ . Let the resulting matrix be  $T$ .

Thus, we are applying the concept of the matrix of a linear mapping to the special case of the identity mapping  $\text{id}_V$ .



The resulting matrix  $T$  is called the *coordinate transition matrix* for changing the basis from  $\underline{u}$  to the basis  $\underline{v}$ .

The fact that the matrix  $T$  of the identity mapping yields exactly the transformation of coordinates between the two bases is easily seen.

Consider the expression of  $u$  with the basis  $\underline{u}$

$$u = x_1u_1 + \dots + x_nu_n,$$

and replace the vectors  $u_i$  by their expressions as linear combinations of the vectors  $v_i$  in the basis  $\underline{v}$ . Collecting the terms properly, we obtain the coordinate expression  $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)$  of the same vector  $u$  in the basis  $\underline{v}$ . It is enough just to reorder the summands and express the individual scalars at the vectors of the basis. But this is exactly what we do when forming the matrix for the identity mapping, thus  $\bar{x} = T \cdot x$ .

We have arrived at the following instruction for building the coordinate transition matrix:

and we should have the relation  $au_1 + bu_2 = cu_3$ . This gives the matrix equation

$$a \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} 4 \\ 3 \\ 5 \end{pmatrix} - c \begin{pmatrix} 5 \\ -10 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

To solve the corresponding homogeneous system, you can use Sage in the following straightforward manner:

```
var("a, b, c")
eq1=2*a+4*b-5*c; eq2=a+3*b+10*c; eq3=5*b-2*c
solve([eq1==0, eq2==0, eq3==0], a, b, c)
```

Sage's output has the form  $[[a == 0, b == 0, c == 0]]$ , and hence  $v = (0, 0, 0)^T$ , i.e.,  $C(A) \cap \text{Ker}(A^T) = \{0\}$ . This proves the first direct sum decomposition and you are encouraged to prove the second one, which is actually easier.  $\square$

*In Mathematics, the structures are best understood via mappings which preserve them. With vector spaces, these are the linear mappings, cf. 2.3.12. We met already homotheties, rotations, and reflections in Chapter 1.*

By the very definition, a linear mapping  $f : V \rightarrow W$  is uniquely determined by its action on the basis vectors of its domain.<sup>9</sup> This results in the unique matrix representation of  $f$ , once the basis on  $V$  and  $W$  are chosen. Thus, we are back at matrix calculus and composition of mappings is given by products of matrices.

Two special subspaces are associated to a linear map  $f : V \rightarrow W$ , the kernel  $\text{Ker}(f) = f^{-1}(\{0\}) \subset V$  and the image  $\text{Im}(f) = f(V) \subset W$ . Isomorphisms are those  $f$  with trivial kernels  $\text{Ker}(f) = \{0\}$  and images  $\text{Im}(f) = W$ . Hence such mappings are both surjective and injective, and their matrices are invertible. Any transformation of coordinates is such an example.

Although we usually write vectors as columns, in the sequel, we shall often write just  $x = (x_1, \dots, x_n)$  for vectors in  $\mathbb{K}^n$ .

**2.C.23.** Check which of  $F$  is a linear transformation:

- i)  $F : \mathbb{R}^3 \rightarrow \mathbb{R}^2, F((x, y, z)) = (y, z)$ ;
- ii)  $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3, F((x, y, z)) = (x + z, 2x - 3 + z, -2y)$ ;
- iii)  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2, F((x, y)) = (x + a, y), 0 \neq a \in \mathbb{R}$ ;
- iv)  $F : \mathbb{R} \rightarrow \mathbb{R}, F(x) = x + a$ , where  $a \in \mathbb{R}$ ;
- v)  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2, F((x, y)) = (x^2, y)$ ;
- vi)  $F : \mathbb{R}^2 \rightarrow \mathbb{R}_1[t], F((x, y)) = y + xt$ ;
- vii)  $F : \text{Mat}_n(\mathbb{R}) \rightarrow \text{Mat}_n(\mathbb{R}), F(A) = AB - BA$ , where  $B \in \text{Mat}_n(\mathbb{R})$  is fixed;
- viii)  $F : \text{Mat}_n(\mathbb{R}) \rightarrow \text{Mat}_n(\mathbb{R}), F(A) = A^T$ ;
- ix)  $F : \mathbb{Z}_3^2 \rightarrow \mathbb{Z}_3^3, F((x, y)) = (x + y, 2x + y, x)$ , where as usual  $\mathbb{Z}_3^n = \mathbb{Z}_3 \times \dots \times \mathbb{Z}_3$  ( $n$ -factors);
- x)  $F : \mathbb{R} \rightarrow \mathbb{R}, F(t) = e^t$ , where  $e$  is the base of the natural logarithm.  $\circ$

<sup>9</sup>Linear mappings  $T : V \rightarrow W$  are also called linear transformations or operators. Linear mappings  $T : V \rightarrow V$  are also called endomorphism of  $V$ .

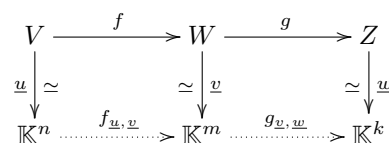
CALCULATING THE MATRIX FOR CHANGING THE BASIS

**Proposition.** The matrix  $T$  for the transition from the basis  $\underline{u}$  to the basis  $\underline{v}$  is obtained by taking the coordinates of the vectors of the basis  $\underline{u}$  expressed in the basis  $\underline{v}$  and writing them as the columns of the matrix  $T$ . The new coordinates  $\bar{x}$  in terms of the new basis  $\underline{v}$  are then  $\bar{x} = T \cdot x$ , where  $x$  is the coordinate vector in the original basis  $\underline{u}$ .

Because the inverse mapping to the identity mapping is again the identity mapping, the coordinate transition matrix is always invertible and its inverse  $T^{-1}$  is the coordinate transition matrix in the opposite direction, that is from the basis  $\underline{v}$  to the basis  $\underline{u}$  (just have a look at the diagram above and invert all the arrows).

**2.3.16. More coordinates.** Next, we are interested in the matrix of a composition of the linear mappings.

Thus, consider another vector space  $Z$  over  $\mathbb{K}$  of dimension  $k$  with basis  $\underline{w}$ , linear mapping  $g : W \rightarrow Z$  and denote the corresponding matrix by  $g_{\underline{v}, \underline{w}}$ .

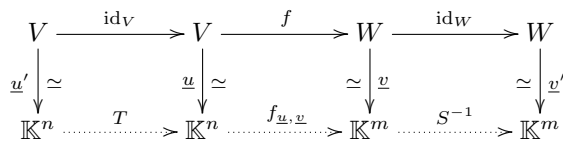


The composition  $g \circ f$  on the upper row corresponds to the matrix of the mapping  $\mathbb{K}^n \rightarrow \mathbb{K}^k$  on the bottom and we calculate directly (we write  $A$  for the matrix of  $f$  and  $B$  for the matrix of  $g$  in the chosen bases):

$$\begin{aligned} g_{\underline{v}, \underline{w}} \circ f_{\underline{u}, \underline{v}}(x) &= \underline{w} \circ g \circ \underline{v}^{-1} \circ \underline{v} \circ f \circ \underline{u}^{-1} \\ &= B \cdot (A \cdot x) = (B \cdot A) \cdot x = (g \circ f)_{\underline{u}, \underline{w}}(x) \end{aligned}$$

for every  $x \in \mathbb{K}^n$ . By the associativity of matrix multiplications, the composition of mappings corresponds to multiplication of the corresponding matrices. Note that the isomorphisms correspond exactly to invertible matrices and that the matrix of the inverse mapping is the inverse matrix.

The same approach shows how the matrix of a linear mapping changes, if we change the coordinates on both the domain and the codomain:



where  $T$  is the coordinate transition matrix from  $\underline{u}'$  to  $\underline{u}$  and  $S$  is the coordinate change matrix from  $\underline{v}'$  to  $\underline{v}$ . If  $A$  is the original matrix of the mapping, then the matrix of the new mapping is given by  $A' = S^{-1}AT$ .

In the special case of a linear mapping  $f : V \rightarrow V$ , that is the domain and the codomain are the same space  $V$ , we express  $f$  usually in terms of a single basis  $\underline{u}$  of the space  $V$ . Then the change from the old basis to the new basis  $\underline{u}'$  with

**2.C.24.** Show that the map  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$  defined by  $T((x, y)) = (x, y, x - y)$  is a linear transformation. Next, find the matrix  $A$  corresponding to  $T$  with respect to the standard bases of  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , respectively.

**Solution.** Let  $u = (u_1, u_2)^T$  and  $v = (v_1, v_2)^T$  be two arbitrary vectors on  $\mathbb{R}^2$ . Then we see that

$$\begin{aligned} T(au + bv) &= T\left(a\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + b\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right) = \\ T\left(\begin{pmatrix} au_1 + bv_1 \\ au_2 + bv_2 \end{pmatrix}\right) &= \begin{pmatrix} au_1 + bv_1 \\ au_2 + bv_2 \\ au_1 + bv_1 - au_2 - bv_2 \end{pmatrix} = \\ a\begin{pmatrix} u_1 \\ u_2 \\ u_1 - u_2 \end{pmatrix} + b\begin{pmatrix} v_1 \\ v_2 \\ v_1 - v_2 \end{pmatrix} &= aT(u) + bT(v), \end{aligned}$$

for all  $a, b \in \mathbb{R}$ . Thus  $T$  is a linear transformation.

Let  $\underline{e} = \{e_1 = (1, 0)^T, e_2 = (0, 1)^T\}$  be the standard basis of  $\mathbb{R}^2$  and denote by  $\underline{\varepsilon} = \{\varepsilon_1 = (1, 0, 0)^T, \varepsilon_2 = (0, 1, 0)^T, \varepsilon_3 = (0, 0, 1)^T\}$  the standard basis of  $\mathbb{R}^3$ . According to the discussion in 2.3.14, the columns of the matrix  $A = (a_{ij})$  corresponding to  $T$  consist of the coordinates of the values  $T(e_j)$ , for  $j = 1, 2$ , the latter expressed in the basis  $\underline{\varepsilon}$  on the target space  $\mathbb{R}^3$ . We compute

$$\begin{aligned} T(e_1) &= (1, 0, 1)^T = 1 \cdot e_1 + 0 \cdot e_2 + 1 \cdot e_3, \\ T(e_2) &= (0, 1, -1)^T = 0 \cdot e_1 + 1 \cdot e_2 - 1 \cdot e_3, \end{aligned}$$

hence the matrix  $A$  has the form  $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{pmatrix}$ . As a simple verification, show that

$$T(u) = Au = (x, y, x - y)^T,$$

for all vectors  $u = (x, y) \in \mathbb{R}^2$ . In fact, the matrix presentation  $T(u) = Au$  provides an easier way to verify the linearity of  $T$ . Indeed, assuming that  $a, b, u, v$  are as above we get  $T(au + bv) = A(au + bv) = a(Au) + b(Av) = aT(u) + bT(v)$ .  $\square$

**2.C.25. Remark.** In Sage we can treat linear mappings via



the command `linear_transformation`, see also the task in 2.C.27 for a complementary application. Let us use the linear mapping  $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$  given in 2.C.24 to illustrate the situation:

```
V=RR^2; W=RR^3
var("x, y"); f(x, y)=[x, y, x-y]
T=linear_transformation(V, W, f)
show(T)
```

Executing this block will display, among other information, the matrix representation of  $T$ , which we present here:

$$\begin{pmatrix} 1.000000000 & 0.000000000 & 1.000000000 \\ 0.000000000 & 1.000000000 & -1.000000000 \end{pmatrix}.$$

Notice this matrix is from the “left”, in terms of Sage, and hence the transpose of the matrix  $A$  presented above.<sup>10</sup> To ensure that Sage will print the correct matrix, add the following code:

<sup>10</sup>This means that  $A$  acts on a vector  $x$  as  $x^T A$ .

the coordinate transition matrix  $T$  leads to the new matrix  $A' = T^{-1}AT$ .

**2.3.17. Linear forms.** A simple but very important case of linear mappings on an arbitrary vector space  $V$  over the scalars  $\mathbb{K}$  appears with the codomain being the scalars themselves, i.e. mappings  $f : V \rightarrow \mathbb{K}$ . We call them *linear forms*.



If we are given the coordinates on  $V$ , the assignments of a single  $i$ -th coordinate to the vectors is an example of a linear form. More precisely, for every choice of basis  $\underline{v} = (v_1, \dots, v_n)$ , there are the linear forms  $v_i^* : V \rightarrow \mathbb{K}$  such that  $v_i^*(v_j) = \delta_{ij}$ , that is,  $v_i^*(v_j) = 1$  when  $i = j$ , and  $v_i^*(v_j) = 0$  when  $i \neq j$ .

The vector space of all linear forms on  $V$  is denoted by  $V^*$  and we call it the *dual space* of the vector space  $V$ . Let us now assume that the vector space  $V$  has finite dimension  $n$ . The basis of  $V^*$ ,  $\underline{v}^* = (v_1^*, \dots, v_n^*)$ , composed of assignments of individual coordinates as above, is called the *dual basis* to  $\underline{v}$ . Clearly this is a basis of the space  $V^*$ , because these forms are evidently linearly independent (prove it!) and if  $\alpha \in V^*$  is an arbitrary form, then for every vector  $u = x_1 v_1 + \dots + x_n v_n$

$$\begin{aligned} \alpha(u) &= x_1 \alpha(v_1) + \dots + x_n \alpha(v_n) \\ &= \alpha(v_1) v_1^*(u) + \dots + \alpha(v_n) v_n^*(u) \end{aligned}$$

and thus the linear form  $\alpha$  is a linear combination of the forms  $v_i^*$ .

Taking into account the standard basis  $\{1\}$  on the one-dimensional space of scalars  $\mathbb{K}$ , any choice of a basis  $\underline{v}$  on  $V$  identifies the linear forms  $\alpha$  with matrices of the type  $1/n$ , that is, with rows  $y$ . The components of these rows are coordinates of the general linear forms  $\alpha$  in the dual basis  $\underline{v}^*$ . Expressing such a form on a vector is then given by multiplying the corresponding row vector  $y$  with the column of the coordinates  $x$  of the vector  $u \in V$  in the basis  $\underline{v}$ :

$$\alpha(u) = y \cdot x = y_1 x_1 + \dots + y_n x_n.$$

Thus we can see that for every finitely dimensional space  $V$ , the dual space  $V^*$  is isomorphic to the space  $V$ . The choice of the dual basis provides such an isomorphism.

In this context we meet again the scalar product of a row of  $n$  scalars with a column of  $n$  scalars. We have worked with it already in the paragraph 2.1.3 on the page 85.

The situation is different for infinitely dimensional spaces. For instance the simplest example of the space of all polynomials  $\mathbb{K}[x]$  in one variable is a vector space with a countable basis with elements  $v_i = x^i$ . As before, we can define linearly independent forms  $v_i^*$ . Every formal infinite sum  $\sum_{i=0}^{\infty} a_i v_i^*$  is now a well-defined linear form on  $\mathbb{K}[x]$ , because it will be evaluated only for a finite linear combination of the basis polynomials  $x^i, i = 0, 1, 2, \dots$

The countable set of all  $v_i^*$  is thus not a basis. Actually, it can be proved that this dual space cannot have a countable basis.



```
A=T.matrix(side="right");show(A)
```

Check yourselves that now Sage prints out the matrix  $A$  posed in 2.C.24.

2.C.26. Let  $T : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be the linear mapping with type

$$T \left( \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right) = \begin{pmatrix} x + y + z \\ x + y + 2z \\ x + y + 3z \end{pmatrix}.$$

Determine its matrix with respect to the standard basis of  $\mathbb{R}^3$ , and next present an answer via Sage. ○

2.C.27. Consider the matrix  $A = \begin{pmatrix} -1 & 2 & 3 \\ 4 & 2 & 0 \end{pmatrix}$ . Find the value  $f(u)$  of the linear mapping  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$  induced by  $A$ , where  $u = (1, 2)^T \in \mathbb{R}^2$ . Next, using Sage, find a basis for the kernel and the image of  $f$ . Is  $f$  injective or surjective?

**Solution.** The function  $f$  has domain  $\mathbb{R}^2$  and target  $\mathbb{R}^3$ , thus its matrix should be  $3 \times 2$ . By our assumption, this means that matrix of  $f$  is the transpose of the given  $A$ . Thus we have

$$f(u) = \begin{pmatrix} -1 & 4 \\ 2 & 2 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 7 \\ 6 \\ 3 \end{pmatrix}.$$

As before, use the command `linear_transformation()` in Sage to verify this result:

```
A = matrix(RR, [[-1, 2, 3], [4, 2, 0]])
f = linear_transformation(A); f
```

This gives:

```
Vector space morphism represented
by the matrix:
[-1.0000000000000000  2.0000000000000000
 3.0000000000000000]
[ 4.0000000000000000  2.0000000000000000
 0.0000000000000000]
Domain: Vector space of dimension 2
over Real Field with 53 bits of precision
Codomain: Vector space of dimension 3
over Real Field with 53 bits of precision
```

We can now evaluate  $f$  at a vector, simply by typing `f([1, 2])`. This prints the vector  $(7, 6, 3)$ . We can also compute the kernel and image of  $f$  by adding the commands

```
f.kernel(); f.image()
```

respectively. For them, the output is

```
Vector space of degree 2 and dimension
0 over Real Field with 53 bits of precision
Basis matrix: []
```

and

```
Vector space of degree 3 and dimension
2 over Rational Field
Basis matrix:
[ 1  0 -3/5]
[ 0  1  6/5]
```

**2.3.18. The length of vectors and scalar product.** When dealing with the geometry of the plane  $\mathbb{R}^2$  in the first chapter we also needed the concept of the length of vectors and their angles, see 1.5.7. For defining these concepts we used the scalar product of two vectors  $u = (x, y)$  and  $v = (x', y')$  in the form  $u \cdot v = xx' + yy'$ .

Indeed, the expression for the length of  $v = (x, y)$  is given by

$$\|v\| = \sqrt{x^2 + y^2} = \sqrt{v \cdot v},$$

while the (oriented) angle  $\varphi$  of two vectors  $u = (x, y)$  and  $v = (x', y')$  is in the planar geometry given by the formula

$$\cos \varphi = \frac{xx' + yy'}{\|v\| \|v'\|}.$$

Note that this scalar product is linear in each of its arguments, and we denote it by  $u \cdot v$  or by  $\langle u, v \rangle$ . The scalar product defined in such a way is symmetric in its arguments and of course  $\|v\| = 0$  if and only if  $v = 0$ . We also see immediately that two vectors in the Euclidean plane are perpendicular whenever their scalar product is zero.

Now we shall mimic this approach for higher dimensions. First, observe that the angle between two vectors is always a two-dimensional concept (we want the angle to be the same in the two-dimensional space containing the two vectors  $u$  and  $v$ ). In the subsequent paragraphs, we shall consider only finitely dimensional vector spaces over real scalars  $\mathbb{R}$ .

#### SCALAR PRODUCT AND ORTHOGONALITY

A *scalar product* on a vector space  $V$  over real numbers is a mapping  $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$  which is symmetric in its arguments, linear in each of them, and such that  $\langle v, v \rangle \geq 0$  and  $\|v\|^2 = \langle v, v \rangle = 0$  if and only if  $v = 0$ .

The number  $\|v\| = \sqrt{\langle v, v \rangle}$  is called the *length*, or *norm*, of the vector  $v$ .

Vectors  $v$  and  $w \in V$  are called *orthogonal* or *perpendicular* whenever  $\langle v, w \rangle = 0$ . We also write  $v \perp w$ . The vector  $v$  is called *normalised* whenever  $\|v\| = 1$ .

The basis of the space  $V$  composed exclusively of mutually orthogonal vectors is called an *orthogonal basis*. If the vectors in such a basis are all normalised, we call the basis *orthonormal*.

A scalar product is very often denoted by the common dot, that is,  $\langle u, v \rangle = u \cdot v$ . Thus, it is then necessary to recognize from the context whether the dot means a product of two vectors (the result is a scalar) or something different (e.g. we often denote the product of matrices and product of scalars in the same way).

Because the scalar product is linear in each of its arguments, it is completely determined by its values on pairs of basis vectors. Indeed, choose a basis  $\underline{u} = (u_1, \dots, u_n)$  of the space  $V$  and denote

$$s_{ij} = \langle u_i, u_j \rangle.$$





respectively. This means that  $\text{Ker}(f) = \{0\}$ , and so  $f$  is injective, but  $\text{Im}(f)$  is 2-dimensional, with a basis determined by the two vectors that Sage indicates in the solution. Hence  $f$  is not surjective. Note that Sage allows us to directly verify the injectivity and surjectivity of  $f$ , as follows:

```
f.is_injective()
f.is_surjective()
```

Sage's output is True and False, respectively.  $\square$

**2.C.28.** Consider  $\mathbb{R}^3$  endowed with its standard basis  $e = \{e_1, e_2, e_3\}$ , and let  $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be a linear mapping satisfying

$$f(e_1) = (0, 1, 2)^T, f(e_2) = (1, 0, 1)^T, f(e_3) = (-1, 1, 1)^T.$$

- (a) Determine the explicit form of  $f$ ;
- (b) Determine the matrix of  $f$  with respect to the standard basis  $e$ ;
- (c) Present bases of the kernel and the image of  $f$ ;
- (d) Find the inverse  $f^{-1}$ , provided it exists.

**Solution.** (a) By assumption we have  $f(e_1) = (0, 1, 2)^T$ ,  $f(e_2) = (1, 0, 1)^T$  and  $f(e_3) = (-1, 1, 1)^T$ . However, any vector  $u = (x, y, z) \in \mathbb{R}^3$  is written as  $u = x \cdot e_1 + y \cdot e_2 + z \cdot e_3$  (here “ $\cdot$ ” is just the scalar multiplication), and since  $f$  is linear we get

$$f(u) = x \cdot f(e_1) + y \cdot f(e_2) + z \cdot f(e_3) = x \cdot (0, 1, 2)^T + y \cdot (1, 0, 1)^T + z \cdot (-1, 1, 1)^T = (y - z, x + z, 2x + y + z)^T,$$

that is,  $f((x, y, z)) = (y - z, x + z, 2x + y + z)^T$ .

(b) We have that

$$f(e_1) = 0 \cdot e_1 + 1 \cdot e_2 + 2 \cdot e_3, \quad f(e_2) = 1 \cdot e_1 + 0 \cdot e_2 + 1 \cdot e_3, \quad f(e_3) = -1 \cdot e_1 + 1 \cdot e_2 + 1 \cdot e_3.$$

Thus, the matrix of  $f$  with respect to the standard basis on  $\mathbb{R}^3$  is given by

$$A = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix}.$$

In Sage we can verify this result by executing the block

```
V=WR*RR**3; var("x, y, z")
f(x, y, z)=[y-z, x+z, 2*x+y+z]
T=linear_transformation(V, W, f)
A=T.matrix(side="right"); show(A)
```

As an alternative, compute first the matrix  $A$ , and then determine the explicit form of  $f$  by the rule  $f(u) = Au$ , for all  $u \in \mathbb{R}^3$ .

(c) Let us compute first a basis of the image  $\text{Im}(f)$  of  $f$ . By the relation  $f(u) = x \cdot f(e_1) + y \cdot f(e_2) + z \cdot f(e_3)$  it is obvious that

$$\text{Im}(f) = \text{span}_{\mathbb{R}}\{f(e_1), f(e_2), f(e_3)\} = \{(0, 1, 2)^T, (1, 0, 1)^T, (-1, 1, 1)^T\}.$$

We should now check the linear independence of these vectors. By elementary row operations on the matrix  $A$  presented above, we obtain

Then from the symmetry of the scalar product we know  $s_{ij} = s_{ji}$  and from the linearity of the product in each of its arguments we get

$$\left\langle \sum_i x_i u_i, \sum_j y_j u_j \right\rangle = \sum_{i,j} x_i y_j \langle u_i, u_j \rangle = \sum_{i,j} s_{ij} x_i y_j.$$

If the basis is orthonormal, the matrix  $S$  is the unit matrix. This proves the following useful claim:

SCALAR PRODUCT IN COORDINATES

**Proposition.** For every orthonormal basis, the scalar product is given by the coordinate expression

$$(1) \quad \langle x, y \rangle = y^T \cdot x.$$

For each basis of the space  $V$  there is the symmetric matrix  $S$  such that the coordinate expression of the scalar product is

$$(2) \quad \langle x, y \rangle = y^T \cdot S \cdot x.$$

Notice, that with symmetric matrix  $S$  it is just a matter of convention in which order we insert the vectors: the formula

$$x^T \cdot S \cdot y = (x^T \cdot S \cdot y)^T = y^T \cdot S^T \cdot x = y^T \cdot S \cdot x$$

produces the same value. However, we shall later consider the second argument as a linear form, thus it seems to be more convenient to use the expression  $y^T \cdot S \cdot x$ .

**2.3.19. Orthogonal complements and projections.** For every fixed subspace  $W \subset V$  in a space with scalar product, we define its *orthogonal complement* as



$$W^\perp = \{u \in V; u \perp v \text{ for all } v \in W\}.$$

It follows directly from the definition that  $W^\perp$  is a vector subspace. If  $W \subset V$  has a basis  $(u_1, \dots, u_k)$  then the description for  $W^\perp$  is given as  $k$  homogeneous equations for  $n$  variables. Thus  $W^\perp$  will have dimension at least  $n - k$ . Also  $u \in W \cap W^\perp$  means that  $\langle u, u \rangle = 0$ , and thus also  $u = 0$  by the definition of scalar product. Clearly then,  $V$  is the direct sum

$$V = W \oplus W^\perp.$$

A linear mapping  $f : V \rightarrow V$  on any vector space is called a *projection*, if we have

$$f \circ f = f.$$

In such a case, we can write, for every vector  $v \in V$ ,

$$v = f(v) + (v - f(v)) \in \text{Im}(f) + \text{Ker}(f) = V$$

and if  $v \in \text{Im}(f)$  and  $f(v) = 0$ , then also  $v = 0$ . Thus the above sum of the subspaces is direct. We say that  $f$  is a projection to the subspace  $W = \text{Im}(f)$  along the subspace  $U = \text{Ker}(f)$ . In words, the projection can be described naturally as follows: we decompose the given vector into a component in  $W$  and a component in  $U$ , and forget the second one.

$$A = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \sim \begin{pmatrix} 2 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix} \sim \begin{pmatrix} 2 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The final matrix is in reduced row echelon form (recall that a quick computation of RREF in Sage occurs by adding in the previous cell the code `A1 = A.rref(); show(A1)`). Therefore, the pivots are lying on the first two columns (use the command `A.pivots()` to verify this). This means that only the vectors  $f(e_1), f(e_2)$  are linearly independent (in fact it is easy to see that  $f(e_3) = f(e_1) - f(e_2)$ ), and so a basis of  $\text{Im}(f)$  has the form  $\{(0, 1, 2)^T, (1, 0, 1)^T\}$ . Thus  $\dim \text{Im}(f) = 2$ , and in particular  $f$  is not surjective. For the kernel, we should have

$$\dim \mathbb{R}^3 = \dim \text{Ker}(f) + \dim \text{Im}(f)$$

from where we get that  $\dim \text{Ker}(f) = 3 - 2 = 1$ . We may use Sage to find a basis of  $\text{Ker}(f)$ , by one of the commands

```
print(A.right_nullity())
A.right_kernel()
```

The first command prints out the dimension of  $\text{Ker}(A)$ , while the output of the second one includes also a basis,

$$\text{Ker}(A) = \text{span}_{\mathbb{R}} \{(1, -1, -1)^T\}.$$

Of course, this also gives the required basis of  $\text{Ker}(f)$ . Indeed, recall the computation of  $\text{Ker}(f)$  relies on solving the homogeneous system  $Au = 0$ , where  $u = (x, y, z)^T \in \mathbb{R}^3$  is an arbitrary vector on  $\mathbb{R}^3$ . This exploits the identification  $\text{Ker}(A) = \text{Ker}(f)$  and gives

$$\begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

or in other words  $\{y - z = 0, x + z = 0, 2x + y + z = 0\}$ . The next step relies on the reduced row echelon form posed above, which first implies that  $z$  is a free variable, say  $z = t \in \mathbb{R}$ . Then we get  $y = z = t \in \mathbb{R}$  and  $x = -z = -t \in \mathbb{R}$ , that is, there are infinitely many solutions of the form  $t \cdot (-1, 1, 1)^T$ , with  $t \in \mathbb{R}$ . Of course, one can replace  $t$  by  $-t$  to obtain the answer that Sage provides.

(d) The given linear transformation is not injective, since we saw that  $\text{Ker}(f) \neq \{0\}$ . Therefore, the inverse  $f^{-1}$  does not exist.  $\square$

**2.C.29.** Consider the endomorphism  $T : \mathbb{C}^3 \rightarrow \mathbb{C}^3$  with  $T(e_1) = (1, 0, i)^T$ ,  $T(e_2) = (0, 1, 1)^T$ ,  $T(e_3) = (i, 1, 0)^T$ , where  $\{e_1, e_2, e_3\}$  denotes the standard basis of  $\mathbb{C}^3$ . Is  $T$  invertible? Describe a solution using Sage, as well.  $\circ$

**2.C.30.** (a) Consider the complex numbers  $\mathbb{C}$  as a real vector space with its standard basis  $\underline{u} = \{1, i\}$ . In this basis determine the matrix of the following linear mappings:

- 1) conjugation,
- 2) multiplication by the number  $(2 + i)$ .

If  $V$  has a scalar product, we say that the projection is orthogonal if the kernel is orthogonal to the image.

Every subspace  $W \neq V$  thus defines an *orthogonal projection* to  $W$ . It is a projection to  $W$  along  $W^\perp$ , given by the unique decomposition of every vector  $u$  into components  $u_W \in W$  and  $u_{W^\perp} \in W^\perp$ , that is, linear mapping which maps  $u_W + u_{W^\perp}$  to  $u_W$ .

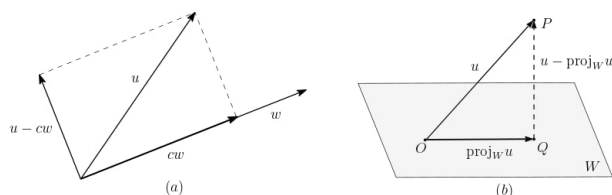
**2.3.20. How to compute projections.** The orthogonal projections  $\text{proj}_v$  to one-dimensional subspaces  $\text{span}\{v\}$  spanned by  $v \in V$  are most simple. Indeed, for any  $u \in V$  we require the projection of  $u$  to be a scalar multiple  $cv$ , such that  $u - cv \perp v$ , i.e.,  $\langle u - cv, v \rangle = 0$  (see part (a) in the figure below). Solving this equation yields the correct value for  $c$  and we deduce

$$\text{proj}_v u = cv = \frac{\langle u, v \rangle}{\langle v, v \rangle} v = \frac{\langle u, v \rangle}{\|v\|^2} v.$$

Consequently, the vector  $u$  decomposes into a pair of orthogonal vectors:  $u = \text{proj}_v u + (u - \text{proj}_v u)$ .

Similarly, for linear subspaces  $W \subset V$  of  $(V, \langle \cdot, \cdot \rangle)$ , there is the direct decomposition  $V = W \oplus W^\perp$  and clearly any vector  $u \in V$  can be uniquely expressed as  $u = w + z$ , where  $w \in W$  and  $z \in W^\perp$ . The vector  $w$  is called the *orthogonal projection of  $u$  in  $W$* , we also write  $\text{proj}_W u$ . The other part of  $u$ , orthogonal to  $W$ , is denoted by  $\text{proj}_{W^\perp} u$  and we have (see part (b) in the figure below):

$$z = \text{proj}_{W^\perp} u = u - \text{proj}_W u.$$



**2.3.21. Existence of orthonormal bases.** It is easy to see that on every finite dimensional real vector space there exist scalar products. Just choose any basis. Define lengths so that each basis vector is of unit length. Call it orthonormal. Immediately we have a scalar product. In this basis the scalar products of vectors are computed as in the formula (1) in the Proposition 2.3.18.



More often we are given a scalar product on a vector space  $V$ , and we want to find an appropriate orthonormal basis for it. We present an algorithm using suitable orthogonal projections in order to transform any basis into an orthogonal one. It is called the *Gramm-Schmidt orthogonalization process*.

The point of this procedure is to transform a given sequence of independent generators  $v_1, \dots, v_k$  of a finite dimensional space  $V$  into an orthogonal set of independent generators of  $V$ .

(b) Determine the matrix of these mappings also in the basis  $\underline{f} = ((1 - i), (1 + i))$ . ○

**2.C.31.** Consider the vector space  $\text{Mat}_{m,n}(\mathbb{K})$  of  $m \times n$  matrices with coefficients in  $\mathbb{K}$ , where  $\mathbb{K}$  is  $\mathbb{R}$  or  $\mathbb{C}$ . Show that  $\dim_{\mathbb{K}} \text{Mat}_{m,n}(\mathbb{K}) = mn$ . Next establish an isomorphism  $\varphi : \text{Mat}_{m,n}(\mathbb{K}) \rightarrow \mathbb{K}^{mn}$ , between  $\text{Mat}_{m,n}(\mathbb{K})$  and  $\mathbb{K}^{mn}$ . ○

**2.C.32.** The vector spaces  $\mathbb{R}_n[x]$  and  $\mathbb{R}^{n+1}$  have the same dimension. Find an isomorphism between them. ○

**2.C.33.** Prove that the vectors

$$u_1 = (1, 2, 0, 0)^T, \quad u_2 = (0, 1, 0, 1)^T, \quad u_3 = (1, 0, 0, 0)^T$$

generate a subspace  $U$  of  $\mathbb{R}^4$  which is isomorphic to  $\mathbb{R}^3$ . Can you find explicitly a linear isomorphism? ○

**2.C.34.** List all subspaces of  $\mathbb{R}^3$ , up to isomorphism. ○

*The simplest linear mappings  $\alpha$  on  $\mathbb{K}^n$  are of the form  $x \mapsto \alpha(x) = c_1x_1 + \dots + c_nx_n \in \mathbb{K}$ , where  $c = (c_1, \dots, c_n)$  is a fixed  $n$ -tuple of scalars from the field  $\mathbb{K}$ , and  $x = (x_1, \dots, x_n)^T \in \mathbb{K}^n$ . Actually, these are all linear mappings  $\mathbb{K}^n \rightarrow \mathbb{K}$ , which can be expressed as  $\alpha(x) = c \cdot x^T$ , through matrix multiplication. Such linear mappings are called linear forms on  $\mathbb{K}^n$ .*

*In fact, any choice of coordinates on a finite-dimensional vector space  $V$  over a field  $\mathbb{K}$ , is actually an  $n$ -tuple of independent linear forms on  $V$ , i.e., linear mappings  $V \rightarrow \mathbb{K}$ . They form the dual vector space  $V^*$  of  $V$ , which is also a vector space over  $\mathbb{K}$ , see 2.3.17. Clearly,  $V$  and  $V^*$  are isomorphic as vector spaces over  $\mathbb{K}$ .*

**2.C.35.** Given the basis  $\underline{u} = \{u_1 = (1, 1)^T, u_2 = (3, 1)^T\}$  of  $\mathbb{R}^2$ , find the dual basis in  $(\mathbb{R}^2)^*$ .

**Solution.** We seek for linear forms  $\varphi_i : \mathbb{R}^2 \rightarrow \mathbb{R}$  satisfying

$$\varphi_1(u_1) = 1, \quad \varphi_1(u_2) = 0, \quad \text{and} \quad \varphi_2(u_1) = 0, \quad \varphi_2(u_2) = 1,$$

respectively. Suppose that  $\varphi_1(x, y) = ax + by$  and  $\varphi_2(x, y) = cx + dy$  for some reals  $a, b, c, d$ . These equations induce the systems  $\{a + b = 1, 3a + b = 0\}$  and  $\{c + d = 0, 3c + d = 1\}$ , respectively. Solving them we deduce that the dual basis to  $\underline{u}$  consists of the linear forms  $\varphi_1(x, y) = -\frac{1}{2}x + \frac{3}{2}y$  and  $\varphi_2(x, y) = \frac{1}{2}x - \frac{1}{2}y$ . □

**2.C.36.** Prove that the set

$$\underline{u} = \{u_1 = (1, 0, 0)^T, u_2 = (0, i, 0)^T, u_3 = (1, 1, i)^T\}$$

is a basis of  $\mathbb{C}^3$ . Next find the dual basis. ○

**2.C.37.** Let  $\{u_1, \dots, u_n\}$  be a basis of a vector space  $V$  over a field  $\mathbb{K}$ , and let  $\{\varphi_1, \dots, \varphi_n\}$  be the dual basis in  $V^*$ . Prove that

(a) Any  $u \in V$  is written as  $u = \sum_{i=1}^n \varphi_i(u)u_i$ .

(b) Any  $\xi \in V^*$  is written as  $\xi = \sum_{i=1}^n \xi(u_i)\varphi_i$ . ○

GRAMM-SCHMIDT ORTHOGONALIZATION

**Proposition.** Let  $(u_1, \dots, u_k)$  be a linearly independent  $k$ -tuple of vectors of a space  $V$  with scalar product. Then there exists an orthogonal system of vectors  $(v_1, \dots, v_k)$  such that  $v_i \in \text{span}\{u_1, \dots, u_i\}$ , and  $\text{span}\{u_1, \dots, u_i\} = \text{span}\{v_1, \dots, v_i\}$ , for all  $i = 1, \dots, k$ . We obtain it by the following procedure:

- The independence of the vectors  $u_i$  ensures that  $u_1 \neq 0$ ; we choose  $v_1 = u_1$ .
- If we have already constructed the vectors  $v_1, \dots, v_\ell$  with the required properties and if  $\ell < k$ , we choose  $v_{\ell+1} = u_{\ell+1} + a_1v_1 + \dots + a_\ell v_\ell$ , where  $a_i = -\frac{\langle u_{\ell+1}, v_i \rangle}{\|v_i\|^2}$ .

**PROOF.** We begin with the first (nonzero) vector  $v_1 = u_1$  and calculate the orthogonal projection  $v_2$  of  $u_2$  to

$$\text{span}\{v_1\}^\perp \subset \text{span}\{v_1, u_2\},$$

i.e., we find the right constant  $a_1$  for which  $v_2 = u_2 + a_1v_1$  is perpendicular to  $v_1$ . The result is nonzero if and only if  $u_2$  is independent of  $v_1$ . All other steps are similar:

In step  $\ell$ ,  $\ell \geq 1$  we seek the vector  $v_{\ell+1} = u_{\ell+1} + a_1v_1 + \dots + a_\ell v_\ell$  satisfying  $\langle v_{\ell+1}, v_i \rangle = 0$  for all  $i = 1, \dots, \ell$ . This implies

$$0 = \langle u_{\ell+1} + a_1v_1 + \dots + a_\ell v_\ell, v_i \rangle = \langle u_{\ell+1}, v_i \rangle + a_i \langle v_i, v_i \rangle$$

and we can see that the vectors with the desired properties are determined uniquely up to a scalar multiple. □

Whenever we have an orthogonal basis of a vector space  $V$ , we just have to normalise the vectors in order to obtain an orthonormal basis. Thus, starting the Gramm-Schmidt orthogonalization with any basis of  $V$ , we have proven:

**Corollary.** On every finite dimensional real vector space with scalar product there exists an orthonormal basis.

In an orthonormal basis, the coordinates and orthogonal projections are very easy to calculate. Indeed, suppose we have an orthonormal basis  $(e_1, \dots, e_n)$  for a space  $V$ . Then every vector  $v = x_1e_1 + \dots + x_n e_n$  satisfies

$$\langle e_i, v \rangle = \langle e_i, x_1e_1 + \dots + x_n e_n \rangle = x_i$$

and so we can always express

$$(1) \quad v = \langle e_1, v \rangle e_1 + \dots + \langle e_n, v \rangle e_n.$$

If we are given a subspace  $W \subset V$  and its orthonormal basis  $(e_1, \dots, e_k)$ , then we can extend it to an orthonormal basis  $(e_1, \dots, e_n)$  for  $V$ . Orthogonal projection of a general vector  $v \in V$  to  $W$  is then given by the expression

$$v \mapsto \langle e_1, v \rangle e_1 + \dots + \langle e_k, v \rangle e_k.$$

In particular, we need only consider an orthonormal basis of the subspace  $W$  in order to write the orthogonal projection to  $W$  explicitly.



Endomorphisms  $F$  on  $V$  are represented by square matrices  $A$  in a given basis. If we leave the choice of the bases free on both domain and target, then we may always achieve the diagonal matrix  $A$  with just ones and zeros on the diagonal, with the rank equal to the dimension of the image, see 2.1.9. But dealing with endomorphisms, we rather want to fix just one basis on  $V$ . Then the available change of the matrix is  $A \mapsto P \cdot A \cdot P^{-1} = B$ , and we call such matrices  $A$  and  $B$  similar. Of course,  $P$  provides the relevant change of basis on  $V$ , see 2.3.16.

Much of our future efforts will aim at understanding linear forms and other concepts within the matrix calculus which are invariant under the above similarity, and thus they reveal properties of the linear mappings themselves. The trace of matrices discussed in the subsequent tasks is an example.

**2.C.38.** Show that the trace  $\text{tr} : \text{Mat}_m(\mathbb{K}) \rightarrow \mathbb{K}$ ,  $\text{tr}(A) = \sum_{i=1}^m a_{ii}$ , is a linear functional, i.e.,  $\text{tr} \in \text{Mat}_m^*(\mathbb{K}) \cong (\mathbb{K}^{m^2})^*$ . Next show that  $\text{tr}(AB) = \text{tr}(BA)$ , for any  $A, B \in \text{Mat}_m(\mathbb{K})$ . Moreover, provide an example verifying that in general  $\text{tr}(AB) \neq \text{tr}(A) \text{tr}(B)$ .  $\circ$

**2.C.39.** Demonstrate with a low-dimensional example that the trace of an endomorphism remains invariant under a change of basis.  $\circ$

**2.C.40.** Prove that if  $A, B$  are two similar square matrices, then  $\text{tr}(A) = \text{tr}(B)$  and  $\det(A) = \det(B)$ . Next decide which of the given pairs of matrices are similar:

(a)  $A = \begin{pmatrix} 1 & -2 \\ 2 & 3 \end{pmatrix}$ ,  $B = \begin{pmatrix} 2 & 2 \\ -1 & 2 \end{pmatrix}$ .

(b)  $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ ,  $B = \begin{pmatrix} \gamma\delta & \delta^2 \\ -\gamma^2 & -\gamma\delta \end{pmatrix}$ ,

where  $\gamma, \delta$  are non-zero real numbers.

(c)  $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ ,  $B = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$ .  $\circ$

**2.C.41. Transition matrix.** Find the transition matrix from the standard basis  $\underline{e} = \{e_1, e_2, e_3\}$  of  $\mathbb{R}^3$  to the basis

$$\underline{u} = \{u_1 = (1, 1, 3)^T, u_2 = (1, -1, 1)^T, u_3 = (3, 1, 5)^T\}.$$

Next, find the coordinates of the vector  $w = (1, 2, 3)^T \in \mathbb{R}^3$  in the new ordered basis  $\underline{u}$ .

**Solution.** We see that

$$\begin{aligned} u_1 &= e_1 + e_2 + 3e_3, \\ u_2 &= e_1 - e_2 + e_3, \\ u_3 &= 3e_1 + e_2 + 5e_3. \end{aligned}$$

Thus, the transition matrix  $T$  from  $\underline{u}$  to  $\underline{e}$  has as columns exactly the vectors of the basis  $\underline{u}$ :

$$T = \begin{pmatrix} 1 & 1 & 3 \\ 1 & -1 & 1 \\ 3 & 1 & 5 \end{pmatrix}.$$

Note that in general the projection  $f$  to the subspace  $W$  along  $U$  and the projection  $g$  to  $U$  along  $W$  is constrained by the equality  $g = \text{id}_V - f$ . Thus, when dealing with orthogonal projections to a given subspace  $W$ , it is always more efficient to calculate the orthonormal basis of that space  $W$  or  $W^\perp$  whose dimension is smaller.

Note also that the existence of an orthonormal basis guarantees that for every real space  $V$  of dimension  $n$  with a scalar product, there exists a linear mapping which is an isomorphism between  $V$  and the space  $\mathbb{R}^n$  with the standard scalar product (i.e. respecting the scalar products as well). We saw already in Theorem 2.3.18 that the desired isomorphism is exactly the coordinate assignment. In words – in every orthonormal basis the scalar product is computed by the same formula as the standard scalar product in  $\mathbb{R}^n$ .

We shall return to the questions of the length of a vector and to projections in the following chapter in a more general context.

**2.3.22. Angle between two vectors.** As we have already noted, the angle between two linearly independent vectors in the space must be the same as when we consider them in the two-dimensional subspace they generate. Basically, this is the reason why the notion of angle is independent of the dimension of the original space. If we choose an orthogonal basis such that its first two vectors generate the same subspace as the two given vectors  $u$  and  $v$  (whose angle we are measuring), we can simply take the definition from the planar geometry. Independently of the choice of coordinates we can formulate the definition as follows:

ANGLE BETWEEN TWO VECTORS

The angle  $\varphi$  between two vectors  $v$  and  $w$  in a vector space with a scalar product is given by the relation

$$\cos \varphi = \frac{\langle v, w \rangle}{\|v\| \|w\|}.$$

The angle defined in this way does not depend on the order of the vectors  $v, w$  and it is chosen in the interval  $0 \leq \varphi \leq \pi$ .

We shall return to scalar products and angles between vectors in further chapters.

**2.3.23. Multilinear forms.** The scalar product was given as a mapping from the product of two copies of a vector space  $V$  into the space of scalars, which was linear in each of its arguments. Similarly, we will work with mappings from the product of  $k$  copies of a vector space  $V$  into the scalars, which are linear in each of its  $k$  arguments. We speak of  $k$ -linear forms.

Most often we will meet *bilinear forms*, that is, the case  $\alpha : V \times V \rightarrow \mathbb{K}$ , where for any four vectors  $u, v, w, z$  and scalars  $a, b, c$  and  $d$  we have

$$\begin{aligned} \alpha(au + bv, cw + dz) &= ac \alpha(u, w) + ad \alpha(u, z) \\ &\quad + bc \alpha(v, w) + bd \alpha(v, z). \end{aligned}$$



The transition matrix from the  $\underline{e}$  to  $\underline{u}$  is now given by the inverse  $T^{-1}$  of  $T$ , see 2.3.15. By inspection, we compute

$$T^{-1} = \begin{pmatrix} -\frac{3}{2} & -\frac{1}{2} & 1 \\ -\frac{1}{2} & -1 & \frac{1}{2} \\ 1 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix},$$

hence the new coordinates of  $w$  are given by the matrix multiplication

$$T^{-1}w = \begin{pmatrix} -\frac{3}{2} & -\frac{1}{2} & 1 \\ -\frac{1}{2} & -1 & \frac{1}{2} \\ 1 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{pmatrix}.$$

□

2.C.42. Solve the problem in 2.C.41 using Sage. ○

2.C.43. Consider the following ordered basis  $\underline{u}_1 = \{E_1 = (1, 0, 1)^T, E_2 = (1, 1, 0)^T, E_3 = (0, 1, 1)^T\}$  of  $\mathbb{R}^3$ , and suppose that the transition matrix  $T$  from another ordered basis  $\underline{u}_2$  of  $\mathbb{R}^3$  to  $\underline{u}_1$  is given by

$$T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}.$$

Find the basis  $\underline{u}_2$ . ○

2.C.44. Suppose that a linear mapping  $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  has the following matrix (with respect to the standard basis)

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ 2 & 0 & 0 \end{pmatrix}.$$

Compute the matrix of this mapping in the new basis  $\underline{f} := \{f_1, f_2, f_3\} := \{(1, 1, 0)^T, (-1, 1, 1)^T, (2, 0, 1)^T\}$ . ○

*In applications, we often know the size of vectors, i.e., there is the so-called notion of norm on the vector space. We saw how to deal with this in Chapter 1 already — the norm is usually defined via the scalar product (see 1.E.20). This is closely linked with the concept of angle between vectors, which is a 2-dimensional concept and works in all dimensions, even the infinite ones, without changes.*

*The scalar product allows us to employ orthogonal projections, orthogonal bases, and other important concepts, see the theoretical column for details. In the following section, we shall come to mappings and matrices compatible with scalar products, which leads to the spectral theory discussed properly in the next chapter. Therefore, consider the remainder of this section as a gentle preparation.*

2.C.45. **Remark.** Consider the vectors  $u = (1, 2, 2)^T$  and  $v = (0, 1, -1)^T$  on  $\mathbb{R}^3$ . Then we see that  $u \cdot v = \langle u, v \rangle = 0$ , that is,  $u, v$  are orthogonal. To encode this we often write  $u \perp v$ . As we know by 1.E.20, in Sage to compute the scalar product  $\langle u, v \rangle = \sum_i u_i v_i$ , of two vectors  $u, v \in \mathbb{R}^n$  we use the command `u.dot_product(v)`. For instance, the block



If additionally we always have

$$\alpha(u, w) = \alpha(w, u),$$

then we speak of a *symmetric bilinear form*. If interchanging the arguments leads to a change of sign, we speak of an *antisymmetric bilinear form*.

Already in planar geometry we have defined the determinant as a bilinear antisymmetric form  $\alpha$ , that is,  $\alpha(u, w) = -\alpha(w, u)$ . In general, due to the theorem 2.2.5, we know that the determinant with dimension  $n$  can be seen as an  $n$ -linear antisymmetric form.

As with linear mappings it is clear that every  $k$ -linear form is completely determined by its values on all  $k$ -tuples of basis elements in a fixed basis. In analogy to linear mappings we can see these values as  $k$ -dimensional analogues to matrices. We show this by an example with  $k = 2$ , where it will correspond to matrices as we have defined them.

#### MATRIX OF A BILINEAR FORM

If we choose a basis  $\underline{u}$  on  $V$  and define for a given bilinear form  $\alpha$  scalars  $a_{ij} = \alpha(u_i, u_j)$  then we obtain for vectors  $v, w$  with coordinates  $x$  and  $y$  (as columns of coordinates)

$$\alpha(v, w) = \sum_{i,j=1}^n a_{ij} x_i y_j = y^T \cdot A \cdot x,$$

where  $A$  is a matrix  $A = (a_{ij})$ .

Directly from the definition of the matrix of a bilinear form we see that the form is symmetric or antisymmetric if and only if the corresponding matrix has this property.

Every bilinear form  $\alpha$  on a vector space  $V$  defines a mapping  $V \rightarrow V^*, v \mapsto \alpha(\cdot, v)$ . That is, by placing a fixed vector in the second argument we obtain a linear form which is the image of this vector. If we choose a fixed basis on a finitely dimensional space  $V$  and a dual basis  $V^*$ , then we have the mapping

$$y \mapsto (x \mapsto y^T \cdot A \cdot x).$$

All this is a matter of convention. Also we may fix the first vector and get a linear form again.

## 4. Properties of linear mappings

In order to exploit vector spaces and linear mappings in modelling real processes and systems in other sciences, we need a more detailed analysis of properties of diverse types of linear mappings.

2.4.1. We begin with four examples in the lowest dimension of interest. With the standard basis of the plane  $\mathbb{R}^2$  and with the standard scalar product we consider the following matrices of mappings  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ :

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, C = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$



```
u = vector(RDF, [1, 2, 2])
v = vector(RDF, [0, 1, -1])
u.dot_product(v)
```

returns the result 0.0, and verifies that  $u \perp v$ . We also know how to compute norms of vectors. For instance, for the given  $u, v$ , we can add the code

```
u.norm(); v.norm()
```

returning 3.0 and  $1.4142135 \approx \sqrt{2}$ , respectively.

Scalar products over complex vector spaces will need differnt approach, we come to them in Chapter 3. They are linked to *Hermitian forms* on  $\mathbb{C}^n$ , and are fundamental in quantum computing.



Let us now turn our attention to vectors in our 3-dimensional world. In the solution to the next task we shall see the general concepts and procedures working independent of the (finite) dimensions. One of the main concepts is that of orthogonal projections, see 2.3.20 for instructions how to compute them.

**2.C.46.** Consider the following four vectors in  $\mathbb{R}^3$ :

- i)  $u_1 = (1, 3, \sqrt{2})^T, u_2 = (-1, 1, -\sqrt{2})^T$ ;
- ii)  $v_1 = (0, 1, 2)^T, v_2 = (-1, 2, 3)^T$ ;

Compute their dot products, norms, angles. Check that the product is symmetric, and next use Sage to solve the tasks.

○

**2.C.47.** Given the linear form  $\alpha : \mathbb{R}^3 \rightarrow \mathbb{R}$  with  $\alpha(u) = 4x + 6y - 2z$  for all  $u = (x, y, z)^T \in \mathbb{R}^3$ , determine the unique vector  $v = (a, b, c)^T \in \mathbb{R}^3$  satisfying  $f(u) = \langle u, v \rangle$ , where  $\langle \cdot, \cdot \rangle$  denotes the usual dot product on  $\mathbb{R}^3$ .

**Solution.** This is an application of the fact that the standard basis  $\{e_1, e_2, e_3\}$  of  $\mathbb{R}^3$  is orthonormal with respect to the dot product, i.e.,  $\langle e_i, e_j \rangle = \delta_{i,j}$  for  $1 \leq i, j \leq 3$ . Then, we have

$$v = \sum_{i=1}^3 \langle v, e_i \rangle e_i = \sum_{i=1}^3 \alpha(e_i) e_i,$$

where we have used that  $\langle \cdot, \cdot \rangle$  is symmetric and moreover that  $\alpha(e_i) = \langle e_i, v \rangle$ , for all  $i$ . Since  $\alpha(e_1) = 4, \alpha(e_2) = 6$  and  $\alpha(e_3) = -2$ , we get  $v = 4e_1 + 6e_2 - 2e_3$ , that is  $v = (4, 6, -2)^T \in \mathbb{R}^3$ . Now one can easily verify that  $\alpha(u) = \langle u, v \rangle$  for all  $u \in \mathbb{R}^3$ . □

**2.C.48. Orthogonal projections via Sage.** It is very easy to compute orthogonal projections of vectors in Sage. For instance, suppose that we want to compute the projection of  $u = (1, 2, 4)^T \in \mathbb{R}^3$  onto  $w = (1, -2, 1)^T \in \mathbb{R}^3$ . We have  $\text{proj}_w(u) = \frac{\langle u, w \rangle}{\|w\|^2} w$ , so it is reasonable to type

```
u = vector ([1 , 2 , 4])
w = vector ([1 , -2 , 1])
proj = u.dot_product (w) / (norm (w) ^2) *w
proj
sage: (1/6, -1/3, 1/6)
```

Hence  $\text{proj}_w(u) = (1/6, -1/3, 1/6)^T \in \mathbb{R}^3$ . As a verification we can check that  $\text{proj}_w(u)$  is orthogonal to  $u - \text{proj}_w(u)$ :

The matrix  $A$  describes the orthogonal projection along the subspace

$$W = \{(0, a); a \in \mathbb{R}\} \subset \mathbb{R}^2$$

to the subspace

$$V = \{(a, 0); a \in \mathbb{R}\} \subset \mathbb{R}^2,$$

that is, the projection to the  $x$ -axis along the  $y$ -axis. Evidently for this  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  we have  $f \circ f = f$  and thus the restriction  $f|_V$  of the given mapping on its codomain is the identity mapping. The kernel of  $f$  is exactly the subspace  $W$ .

The matrix  $B$  has the property  $B^2 = 0$ , therefore the same holds for the corresponding mapping  $f$ . We can envision this as the differentiation of polynomials  $\mathbb{R}_1[x]$  of degree at most one in the basis  $(1, x)$  (we shall come to differentiation in chapter five, see 5.1.6).

The matrix  $C$  gives a mapping  $f$ , which rescales the first vector of the basis  $a$ -times, and the second one  $b$ -times. Therefore the whole plane divides into two subspaces, which are preserved under the mapping and where it is only a *homothety*, that is, scaling by a scalar multiple (the first case was a special case with  $a = 1, b = 0$ ). For instance the choice  $a = 1, b = -1$  corresponds to axial symmetry (mirror symmetry) under the  $x$ -axis, which is the same as complex conjugation  $x+iy \mapsto x-iy$  on the two-dimensional real space  $\mathbb{R}^2 \simeq \mathbb{C}$  in basis  $(1, i)$ . This is a linear mapping of the two-dimensional real vector space  $\mathbb{C}$ , but not of the one-dimensional complex space  $\mathbb{C}$ .

The matrix  $D$  is the matrix of rotation by 90 degrees (the angle  $\pi/2$ ) centered at the origin in the standard basis. We can see at first glance that none of the one-dimensional subspaces is preserved under this mapping.

Such a rotation is a bijection of the plane onto itself, therefore we can surely find distinct bases in the domain and codomain, where its matrix will be the unit matrix  $E$ . We simply take any basis of the domain and its image in the codomain. But we are not able to do this with the same basis for both the domain and the codomain.

Consider the matrix  $D$  as a matrix of the mapping  $g : \mathbb{C}^2 \rightarrow \mathbb{C}^2$  with the standard basis of the complex vector space  $\mathbb{C}^2$ . Then we can find vectors  $u = (i, 1), v = (-i, 1)$ , for which we have



$$g(u) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ i \end{pmatrix} = i \cdot u,$$

$$g(v) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} -i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -i \end{pmatrix} = -i \cdot v.$$

That means that in the basis  $(u, v)$  on  $\mathbb{C}^2$ , the mapping  $g$  has the matrix

$$K = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Notice that by extending the scalars to  $\mathbb{C}$ , we arrive at an analogy to the matrix  $C$  with diagonal elements  $a = \cos(\frac{1}{2}\pi) + i \sin(\frac{1}{2}\pi)$  and its complex conjugate  $\bar{a}$ . In other words, the

```
proj.dot_product(u-proj)
```

which gives 0.

Recall now that a nice feature of Sage is that we can easily extend its capabilities by introducing new commands. Here will create a function that computes the orthogonal projection of a vector onto another, a procedure which allows us to view (in Sage) the orthogonal projection as a function of two vectors. For this, use the cell

```
def proj (u,w):
    a = u. dot_product (w)/(norm (w)^2)*w
    return(a)
```

which defines the projection of  $u$  onto  $w$ . In such terms, the following cell in Sage gives the same result as above.

```
u1 = vector ([1 , 2 , 4])
u2 = vector ([1 , -2 , 1])
def proj (u,w):
    a = u. dot_product (w)/(norm (w)^2)*w
    return(a)
proj (u1 ,u2)
sage: (1/6, -1/3, 1/6)
```

Try implementing this block on your own. The function defined here will be useful for implementing the Gram-Schmidt procedure, which we are going to describe very soon.

**2.C.49.** Determine the matrix  $A$  which, under the standard basis of  $\mathbb{R}^3$ , gives the orthogonal projection on the vector subspace generated by the vectors  $u_1 = (-1, 1, 0)^T$  and  $u_2 = (-1, 0, 1)^T$ .

**Solution.** First observe that given subspace is a plane containing the origin of  $\mathbb{R}^3$ , with normal vector  $u_3 = (1, 1, 1)^T$ . This is because the ordered triple  $(1, 1, 1)$  is clearly a solution to the system  $\{-x_1 + x_2 = 0, -x_1 + x_3 = 0\}$ , that is, the vector  $u_3$  is perpendicular to the vectors  $u_1, u_2$ . Under the given projection the vectors  $u_1$  and  $u_2$  must map to themselves, and the vector  $u_3$  to the zero vector. Thus, with respect to the basis  $\{u_1, u_2, u_3\}$ , the matrix presentation of the projection is

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

If the transition matrix from the basis  $\{u_1, u_2, u_3\}$  to the standard basis is denoted by  $T$ , then the transition matrix from the standard basis to  $\{u_1, u_2, u_3\}$  is the inverse  $T^{-1}$ . We compute

$$T = \begin{pmatrix} -1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix},$$

and thus we obtain

$$A = TPT^{-1} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

argument of the number  $a$  in polar form provides the angle of the rotation.

This is easy to understand, if we denote the real and imaginary part of the vector  $u$  as follows

$$u = x_u + iy_u = \operatorname{Re} u + i \operatorname{Im} u = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + i \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The vector  $v$  is the complex conjugate of  $u$ . We are interested in the restriction of the mapping  $g$  to the real vector subspace  $V = \mathbb{R}^2 \cap \operatorname{span}_{\mathbb{C}}\{u, v\} \subset \mathbb{C}^2$ . Evidently,

$$V = \operatorname{span}_{\mathbb{R}}\{u + \bar{u}, i(u - \bar{u})\} = \operatorname{span}_{\mathbb{R}}\{x_u, -y_u\}$$

is the whole plane  $\mathbb{R}^2$ . The restriction of  $g$  to this plane is exactly the original mapping given by the matrix  $D$  (notice this matrix is real, thus it preserves this real subspace). It is immediately seen that this is the rotation through the angle  $\frac{1}{2}\pi$  in the positive sense with respect to the chosen basis  $x_u, -y_u$ . Work it by yourself with a direct calculation. Note also why exchanging the order of the vectors  $u$  and  $v$  leads to the same result, although in a different real basis!

**2.4.2. Eigenvalues and eigenvectors of mappings.** A key



to the description of mappings in the previous examples was the answer to the question “what are the vectors satisfying the equation  $f(u) = a \cdot u$  for some suitable scalars  $a$ ?”.

We consider this question for any linear mapping  $f : V \rightarrow V$  on a vector space of dimension  $n$  over scalars  $\mathbb{K}$ . If we imagine such an equality written in coordinates, i.e. using the matrix of the mapping  $A$  in some bases, we obtain a system of linear equations

$$A \cdot x - a \cdot x = (A - a \cdot E) \cdot x = 0$$

with an unknown parameter  $a$ . We know already that such a system of equations has only the solution  $x = 0$  if the matrix  $A - aE$  is invertible. Thus we want to find such values  $a \in \mathbb{K}$  for which  $A - aE$  is not invertible, and for that, the necessary and sufficient condition reads (see Theorem 2.2.11)

$$(1) \quad \det(A - a \cdot E) = 0.$$

If we consider  $\lambda = a$  as a variable in the previous scalar equation, we are actually looking for the roots of a polynomial of degree  $n$ . As we have seen in the case of the matrix  $D$ , the roots may exist in an extension of our field of scalars, if they are not in  $\mathbb{K}$ .

EIGENVALUES AND EIGENVECTORS

Scalars  $\lambda \in \mathbb{K}$  satisfying the equation  $f(u) = \lambda \cdot u$  for some nonzero vector  $u \in V$  are called the *eigenvalues of the mapping  $f$* . The corresponding nonzero vectors  $u$  are called the *eigenvectors of the mapping  $f$* .

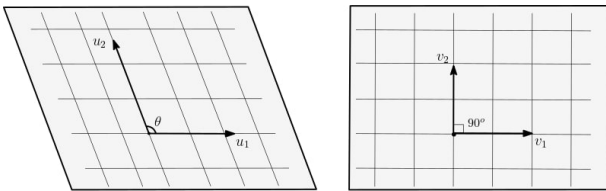
If  $u, v$  are eigenvectors associated with the same eigenvalue  $\lambda$ , then for every linear combination of  $u$  and  $v$ ,

$$f(au + bv) = af(u) + bf(v) = \lambda(au + bv).$$

Therefore the eigenvectors associated with the same eigenvalue  $\lambda$ , together with the zero vector, form a nontrivial vector

Our choice of basis of a given (finite dimensional) vector space corresponds to a coordinate system on it (we get the basis with fixed ordering of its elements). If the basis  $\underline{u} = \{u_1, \dots, u_m\}$  is orthonormal, i.e.  $\langle u_i, u_j \rangle = \delta_{ij}$ , the Kronecker delta, for all  $i, j$ , then the coordinates are given just by the scalar products, i.e.,  $v = \langle v, u_i \rangle u_i + \dots + \langle v, u_m \rangle u_m$  for all  $v \in V$ . With orthogonal basis, we just request that its elements are mutually perpendicular and in the latter expression, we have to divide the scalar product by the norm of the elements, i.e., the coordinates are  $\frac{\langle v, u_i \rangle}{\|u_i\|}$ .

The figure below compares a general basis and an orthonormal basis in  $\mathbb{R}^2$ . Notice that the formula for the scalar products is the same in all orthonormal bases, just the standard Euclidean dot product, as we know it from the matrix calculus. Orthogonal or orthonormal bases are easily obtained by the Gram-Schmidt procedure, see 2.3.21.



**2.C.50.** (a) Prove that the vectors  $E_1 = (1, 1, 1, 1)^T$ ,  $E_2 = (1, 1, 1, 0)^T$  and  $E_3 = (1, 1, 0, 0)^T$  of  $\mathbb{R}^4$  are linearly independent.

(b) Next find an orthonormal basis of the 3-dimensional space  $W = \text{span}_{\mathbb{R}}\{E_1, E_2, E_3\}$  of  $\mathbb{R}^4$ .

(c) Find the new coordinates of the vector  $w = (4, 4, 2, 1)^T \in W$  with respect to the orthonormal basis that you described in part (b).

**Solution.** (a) We check the linear independence by Sage

```
V = RR^4
E1 = vector(RR, [1, 1, 1, 1])
E2 = vector(RR, [1, 1, 1, 0])
E3 = vector(RR, [1, 1, 0, 0])
V.linear_dependence([E1, E2, E3]) == []
```

Sage's output is indeed True.

(b) Let us apply the Gram-Schmidt method to obtain an orthonormal basis of  $W$ . Set  $w_1 = E_1$  and

$$\begin{aligned} w_2 &= E_2 - \text{proj}_{w_1}(E_2) \\ &= E_2 - \frac{\langle E_2, w_1 \rangle}{\|w_1\|^2} w_1, \\ w_3 &= E_3 - \text{proj}_{w_2}(E_3) - \text{proj}_{w_1}(E_3) \\ &= E_3 - \frac{\langle E_3, w_2 \rangle}{\|w_2\|^2} w_2 - \frac{\langle E_3, w_1 \rangle}{\|w_1\|^2} w_1. \end{aligned}$$

We compute  $\|w_1\|^2 = 4$  and  $\langle E_2, E_1 \rangle = 3$ , hence  $w_2 = (1/4, 1/4, 1/4, -3/4)^T$  with  $\|w_2\|^2 = 3/4$ . We proceed by computing  $\langle E_3, w_2 \rangle = 1/2$ ,  $\langle E_3, w_1 \rangle = 2$  and hence

subspace  $V_\lambda \subset V$ . We call it the *eigenspace associated with*  $\lambda$ . For instance, if  $\lambda = 0$  is an eigenvalue, the kernel  $\text{Ker } f$  is the eigenspace  $V_0$ .

We have seen how to compute the eigenvalues in coordinates. The independence of the eigenvalues from the choice of coordinates is clear from their definition. But let us look explicitly what happens if we change the basis. As a direct corollary of the transformation properties from the paragraph 2.3.16 and the Cauchy theorem 2.2.7 for calculation of the determinant of product, the matrix  $A'$  in the new coordinates will be  $A' = P^{-1}AP$  with an invertible matrix  $P$ . Thus

$$\begin{aligned} |P^{-1}AP - \lambda E| &= |P^{-1}AP - P^{-1}\lambda EP| \\ &= |P^{-1}(A - \lambda E)P| \\ &= |P^{-1}||A - \lambda E||P| \\ &= |A - \lambda E|, \end{aligned}$$

because the scalar multiplication is commutative and we know that  $|P^{-1}| = |P|^{-1}$ .

For these reasons we use the same terminology for matrices and mappings:

CHARACTERISTIC POLYNOMIALS

For a matrix  $A$  of dimension  $n$  over  $\mathbb{K}$  we call the polynomial  $|A - \lambda E| \in \mathbb{K}_n[\lambda]$  the *characteristic polynomial of the matrix*  $A$ .

Roots of this polynomial are the *eigenvalues of the matrix*  $A$ . If  $A$  is the matrix of the mapping  $f : V \rightarrow V$  in a certain basis, then  $|A - \lambda E|$  is also called the *characteristic polynomial of the mapping*  $f$ .

Because the characteristic polynomial of a linear mapping  $f : V \rightarrow V$  is independent of the choice of the basis of  $V$ , the coefficients of individual powers of the variable  $\lambda$  are scalars expressing some properties of  $f$ . In particular, they too cannot depend on the choice of the basis. Suppose  $\dim V = n$  and  $A = (a_{ij})$  is the matrix of the mapping in some basis. Then

$$\begin{aligned} |A - \lambda \cdot E| &= (-1)^n \lambda^n + (-1)^{n-1} (a_{11} + \dots + a_{nn}) \lambda^{n-1} \\ &\quad + \dots + |A| \lambda^0. \end{aligned}$$

The coefficient at the highest power says whether the dimension of the space  $V$  is even or odd.

The most interesting coefficient is the sum of the diagonal elements of the matrix. We have just proved that it does not depend on the choice of the basis and we call it the *trace of the matrix*  $A$  and denote it by  $\text{Tr } A$ . The *trace of the mapping*  $f$  is defined as a trace of the matrix in an arbitrary basis.

In fact, this is not so surprising once we notice that the trace is actually the linear approximation of the determinant in the neighbourhood of the unit matrix in the direction  $A$ . We shall deal with such concepts in Chapter 8 only. But since the determinant is a polynomial, we may see easily that the only terms in  $\det(E + tA)$  which are linear in the real parameter  $t$



$w_3 = (1/3, 1/3, -2/3, 0)^T$ . The orthogonality is verified easily:  $\langle w_1, w_2 \rangle = \langle w_1, w_3 \rangle = \langle w_2, w_3 \rangle = 0$ . We also compute  $\|w_3\|^2 = 2/3$ , hence the orthonormal basis  $\{\hat{w}_i = \frac{w_i}{\|w_i\|} : i = 1, 2, 3\}$  of  $W$  has the explicit form

$$\hat{w}_1 = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix}, \hat{w}_2 = \begin{pmatrix} \sqrt{3}/6 \\ \sqrt{3}/6 \\ \sqrt{3}/6 \\ -\sqrt{3}/2 \end{pmatrix}, \hat{w}_3 = \begin{pmatrix} \sqrt{6}/6 \\ \sqrt{6}/6 \\ -\sqrt{6}/3 \\ 0 \end{pmatrix}.$$

(c) The given vector  $w = (4, 4, 2, 1)^T$  lies on  $W$ , since we have the expression  $w = E_1 + E_2 + 2E_3$ . To find its coordinates with respect to the orthonormal basis  $\{\hat{w}_i : i = 1, 2, 3\}$  of  $W$  constructed above, we apply the formula

$$w = \sum_{i=1}^3 \langle w, \hat{w}_i \rangle \hat{w}_i.$$

We compute  $\langle w, \hat{w}_1 \rangle = 11/2$ ,  $\langle w, \hat{w}_2 \rangle = 7\sqrt{3}/6$  and  $\langle w, \hat{w}_3 \rangle = 2\sqrt{6}/3$ , thus  $w = \frac{11}{2}\hat{w}_1 + \frac{7\sqrt{3}}{6}\hat{w}_2 + \frac{2\sqrt{6}}{3}\hat{w}_3$ . Hence the initial coordinates  $(1, 1, 2)$  of  $w$  with respect to the basis  $\{E_1, E_2, E_3\}$  changed to the new coordinates  $(11/2, 7\sqrt{3}/6, 2\sqrt{6}/3)$  with respect to the basis  $\{\hat{w}_1, \hat{w}_2, \hat{w}_3\}$ .  $\square$

**2.C.51.** Use Sage via the function `proj(v, u)` introduced in 2.C.48 to confirm the expression of the orthonormal basis  $\{\hat{w}_1, \hat{w}_2, \hat{w}_3\}$  presented in 2.C.50.  $\circ$

**2.C.52.** Apply the Gram-Schmidt orthogonalisation process to obtain the orthogonal basis of the linear subspace  $U \subset \mathbb{R}^4$

$$U = \{(x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4; x_1 + x_2 + x_3 + x_4 = 0\} \circ$$

**2.C.53.** Consider the linear mapping  $\varphi : \mathbb{R}^4 \rightarrow \mathbb{R}^4$  whose matrix with respect to the standard basis of  $\mathbb{R}^4$  is given by

$$A = \begin{pmatrix} 1/2 & 2 & -1/2 & -1/2 \\ 1 & 1/2 & 1 & 3/2 \\ 2 & 9/2 & 0 & 1/2 \\ 2 & 2 & 0 & 0 \end{pmatrix}.$$

Find an orthonormal basis of the kernel of  $\varphi$ .  $\circ$

**2.C.54.** Consider the Euclidean space  $\mathbb{R}^4$  with its standard dot product  $\langle \cdot, \cdot \rangle$ . Find a basis of the linear subspace of  $\mathbb{R}^4$  consisting of vectors which are orthogonal to:

- i)  $u = (1, 1, 1, 1)^T \in \mathbb{R}^4$ ;
- ii)  $w = (1, 0, 0, -1)^T$  and  $z = (-1, 0, 1, 0)^T \in \mathbb{R}^4$ .  $\circ$

**2.C.55. Orthogonal complement.** Let  $W$  be the subspace of  $\mathbb{R}^4$  spanned by the vectors  $u_1 = (-1, 1, 3, 0)^T$  and  $u_2 = (0, 0, 0, 1)^T$ . Find the orthogonal complement of  $W$ .

**Solution.** Any vector  $x \in W$  is written as  $x = au_1 + bu_2 = (-a, a, 3a, b)^T$  for some  $a, b \in \mathbb{R}$ . Vectors  $y \in W^\perp$  should satisfy  $x \cdot y = 0$ , i.e.,  $-ay_1 + ay_2 + 3ay_3 + by_4 = 0$ , for any  $a, b \in \mathbb{R}$ . For  $a = 1, b = 0$  (this corresponds to the condition  $u_1 \cdot y = 0$ ) and  $a = 0, b = 1$  (this corresponds to the condition  $u_2 \cdot y = 0$ ) we obtain the following system of equations:  $\{-y_1 + y_2 + 3y_3 = 0, y_4 = 0\}$ . Viewing  $y_2 = c \in \mathbb{R}$  and  $y_3 = d \in \mathbb{R}$  as free variables, we obtain the solution

are just the trace. We shall see relation to matrix exponential later in Chapter 8.

The coefficient at  $\lambda^0$  is the determinant  $|A|$  and we shall see later that it describes the rescaling of volumes by the mapping.

**2.4.3. Basis of eigenvectors.** We discuss a few important properties of eigenspaces now.

**Theorem.** *Eigenvectors of linear mappings  $f : V \rightarrow V$  associated to different eigenvalues are linearly independent.*

**PROOF.** Let  $a_1, \dots, a_k$  be distinct eigenvalues of the mapping  $f$  and  $u_1, \dots, u_k$  eigenvectors with these eigenvalues. The proof is by induction on the number of linearly independent vectors among the chosen ones.

Assume that  $u_1, \dots, u_\ell$  are linearly independent and  $u_{\ell+1} = \sum_i c_i u_i$  is their linear combination. We can choose  $\ell = 1$ , because the eigenvectors are nonzero. But then  $f(u_{\ell+1}) = a_{\ell+1} \cdot u_{\ell+1} = \sum_{i=1}^\ell a_{\ell+1} \cdot c_i \cdot u_i$ , that is,

$$f(u_{\ell+1}) = \sum_{i=1}^\ell a_{\ell+1} \cdot c_i \cdot u_i = \sum_{i=1}^\ell c_i \cdot f(u_i) = \sum_{i=1}^\ell c_i \cdot a_i \cdot u_i.$$

By subtracting the second and the fourth expression in the equalities we obtain  $0 = \sum_{i=1}^\ell (a_{\ell+1} - a_i) \cdot c_i \cdot u_i$ . All the differences between the eigenvalues are nonzero and at least one coefficient  $c_i$  is nonzero. This is a contradiction with the assumed linear independence  $u_1, \dots, u_\ell$ , therefore also the vector  $u_{\ell+1}$  must be linearly independent of the others.  $\square$

The latter theorem can be seen as a decomposition of a linear mapping  $f$  into a sum of much simpler mappings. If there are  $n = \dim V$  distinct eigenvalues  $\lambda_i$ , we obtain the entire  $V$  as a direct sum of one-dimensional eigenspaces  $V_{\lambda_i}$ . Each of them then describes a projection on this invariant one-dimensional subspace, where the mapping is given just as multiplication by the eigenvalue  $\lambda_i$ .

Furthermore, this decomposition can be easily calculated:

$(c + 3d, c, d, 0)$ . It follows that  $W^\perp = \text{span}_{\mathbb{R}}\{w_1, w_2\}$ , where the vectors  $w_1, w_2$  are given by  $w_1 = (1, 1, 0, 0)^T$  and  $w_2 = (3, 0, 1, 0)^T$ , respectively.  $\square$

**2.C.56.** Given the dot products on  $\mathbb{R}^m$  and  $\mathbb{R}^n$ , prove that for any  $m \times n$  real matrix  $A$  the column space  $C(A)$  is orthogonal to the left null space  $\text{Ker}(A^T)$  and the row space  $C(A^T)$  is orthogonal to the kernel  $\text{Ker}(A)$  of  $A$ . Next illustrate these statements using the matrix  $A$  described in 2.C.20.  $\circ$

**D. Properties of linear mappings**

A central attention in numerical matrix calculus is devoted to the simplest possible representation of linear mappings, i.e., the optimal choice of coordinates allowing to understand the mappings best. The most perfect understanding of  $f : V \rightarrow V$  comes if we have the mapping represented by a diagonal matrix. This leads to the concepts of “eigenvectors” and “eigenvalues” of  $f$ .



Thus, an eigenvector of an  $n \times n$  matrix  $A$  is a non-trivial solution  $x$  to the equation  $Ax = \lambda x$  with an unknown scalar  $\lambda$ . Obviously, this homogeneous system of equations has got a non-trivial solution if and only if  $\lambda$  is a root of the characteristic polynomial  $\chi_A(\lambda) := \det(A - \lambda E)$ , where  $E$  is the  $n \times n$  identity matrix. We talk about eigenvalues of  $A$ , they may come with multiplicities, and the spectrum of  $A$  consists of the eigenvalues of  $A$ , including their multiplicities (called algebraic multiplicities).

By the very definition, all this corresponds to  $f(u) = \lambda u$ , i.e., these concepts are independent of the choice of coordinates. If we find enough eigenvalues and eigenvectors, the mapping  $f$  enjoys a diagonal matrix!

This may fail for two reasons: the polynomial might not have enough roots in the chosen field  $\mathbb{K}$ , or the number of independent eigenvectors to a given eigenvalue  $\lambda$  might be smaller than the algebraic multiplicity of  $\lambda$ , i.e., the dimension of the eigen-space  $V_\lambda$  spanned by all such eigenvectors is too small (we call it the geometric multiplicity). We met both examples in the first chapter in the plane (rotations and derivatives of first order polynomials, see also the task presented in 2.E.57).

**2.D.1. Eigenvalues and eigenvectors.** Find the eigenvalues and the eigenvectors of the matrix

$$A = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 3 & 0 \\ 2 & -2 & 2 \end{pmatrix}.$$

**Solution.** We begin with the characteristic polynomial of  $A$ :

$$\chi_A(\lambda) = \begin{vmatrix} -1 - \lambda & 1 & 0 \\ -1 & 3 - \lambda & 0 \\ 2 & -2 & 2 - \lambda \end{vmatrix} = \lambda^3 - 4\lambda^2 + 2\lambda + 4.$$

The eigenvalues of  $A$  are the roots of  $\chi_A(\lambda)$ . Using for example the Horner’s scheme presented in Chapter 1, we compute

BASIS OF EIGENVECTORS

**Corollary.** If there exist  $n$  mutually distinct roots  $\lambda_i$  of the characteristic polynomial of the mapping  $f : V \rightarrow V$  on the  $n$ -dimensional space  $V$ , then there exists a decomposition of  $V$  into a direct sum of eigenspaces each of dimension one. This means that there exists a basis for  $V$  consisting only of eigenvectors and in this basis the matrix for  $f$  is the diagonal matrix with the eigenvalues on the diagonal. This basis is uniquely determined up to the order of the elements and scale of the vectors.

The corresponding basis (expressed in the coordinates in an arbitrary basis of  $V$ ) is obtained by solving  $n$  systems of homogeneous linear equations of  $n$  variables with matrices  $(A - \lambda_i \cdot E)$ , where  $A$  is the matrix of  $f$  in a chosen basis.

**2.4.4. Invariant subspaces.** We have seen that every eigenvector  $v$  of the mapping  $f : V \rightarrow V$  generates a subspace  $\text{span}\{v\} \subset V$ , which is preserved by the mapping  $f$ .



More generally, we say that a vector subspace  $W \subset V$  is an invariant subspace for a linear mapping  $f$ , if  $f(W) \subset W$ .

If  $V$  is a finite dimensional vector space and we choose some basis  $(u_1, \dots, u_k)$  of a subspace  $W$ , we can always extend it to a basis  $(u_1, \dots, u_k, u_{k+1}, \dots, u_n)$  for the whole space  $V$ . For every such basis,  $f(W) \subset W$  implies the mapping  $f$  will have the matrix  $A$  of the form

$$(1) \quad A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$$

where  $B$  is a square matrix of dimension  $k$ ,  $D$  is a square matrix of dimension  $n - k$  and  $C$  is a matrix of the type  $n/(n - k)$ . On the other hand, if for some basis  $(u_1, \dots, u_n)$  the matrix of the mapping  $f$  is of the form (1), then  $W = \text{span}\{u_1, \dots, u_k\}$  is invariant under the mapping  $f$ .

By the same arguments, the mapping with the matrix  $A$  as in (1) leaves the subspace  $\text{span}\{u_{k+1}, \dots, u_n\}$  invariant, if and only if the submatrix  $C$  is zero.

From this point of view the eigenspaces of the mapping are special cases of invariant subspaces. Our next task is to find some conditions under which there are invariant complements of invariant subspaces.

**2.4.5.** We illustrate some typical properties of mappings on the spaces  $\mathbb{R}^3$  and  $\mathbb{R}^2$  in terms of eigenvalues and eigenvectors.

(1) Consider the mapping given in the standard basis by the matrix  $A$

$$f : \mathbb{R}^3 \rightarrow \mathbb{R}^3, A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

$\lambda_1 = 2$ ,  $\lambda_{\pm} = 1 \pm \sqrt{3}$ , all with multiplicity one<sup>11</sup> Hence the algebraic multiplicities of  $\lambda_1$  and  $\lambda_{\pm}$  are all one. It follows that each has a 1-dimensional eigenspace, i.e., the geometric multiplicity of each eigenvalue is also one (see 3.4.10).

The eigenvector associated to  $\lambda_1$  is determined by solving the matrix equation  $(A - \lambda_1 E)x = 0$ , with  $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ . This reduces to the system

$$\{-3x_1 + x_2 = 0, -x_1 + x_2 = 0, x_1 - x_2 = 0\}$$

and in order to find a solution we may apply what we learned in Section A. We see that  $x_3$  is a free variable,  $x_3 \in \mathbb{R}$ , and the rest two variables must satisfy  $x_1 = x_2 = 0$ .<sup>12</sup> Thus, the eigenvector associated to  $\lambda_1$  is the vector  $u_1 := (0, 0, 1)^T$  (or any multiple of it), and in particular  $V_{\lambda_1}$  is 1-dimensional (spanned by  $u_1$ ).

Similarly, the eigenvector corresponding to  $\lambda_+ = 1 + \sqrt{3}$  arises by solving the matrix equation  $(A - \lambda_+ E)x = 0$ . This gives the system

$$\begin{cases} -(2 + \sqrt{3})x_1 + x_2 = 0, \\ -x_1 + (2 - \sqrt{3})x_2 = 0, \\ 2x_1 - 2x_2 + (1 - \sqrt{3})x_3 = 0, \end{cases}$$

whose solution has the form  $\{(2 - \sqrt{3}, 1, -2)t : t \in \mathbb{R}\}$ . This implies that the eigenspace  $V_{\lambda_+}$  corresponding to  $\lambda_+$  is also 1-dimensional, as well:

$$V_{\lambda_+} = \text{span}_{\mathbb{R}} \left\{ (2 - \sqrt{3}, 1, -2)^T \right\}.$$

In an analogous way we obtain that the eigenspace associated to the third eigenvalue  $\lambda_-$  is 1-dimensional,

$$V_{\lambda_-} = \text{span}_{\mathbb{R}} \left\{ (2 + \sqrt{3}, 1, -2)^T \right\}. \quad \square$$

**2.D.2. Eigentheory via Sage.** In Sage, given a squared matrix it is easy to compute its eigenvalues and eigenvectors. Let us use the matrix  $A$  from the previous task to illustrate the situation. To introduce the characteristic polynomial and find its roots, we can type

```
A=matrix(SR, [[-1, 1, 0], [-1, 3, 0],
[2, -2, 2]])
p(t) = A.characteristic_polynomial(t)
show(p(t)); show(p.roots())
```

The output here is the characteristic polynomial (with respect to the variable  $t$ ), i.e.,  $t^3 - 4t^2 + 2t + 4$ , and its roots, presented in a list as follows:

$$\left[ (-\sqrt{3} + 1, 1), (\sqrt{3} + 1, 1), (2, 1) \right].$$

<sup>11</sup>Use also Sage to solve the equation  $\chi_A(\lambda) = 0$ , and then read below for a variety of techniques in Sage for finding the eigenvalues of a given matrix  $A$ .

<sup>12</sup>Although this case is extremely easy and the fact that  $x_3$  is a free variable is obvious (as the solution), recall that a good technique solving the homogeneous system  $(A - \lambda E)x = 0$ , where  $\lambda$  is an eigenvalue of  $A$ , relies on the description of the reduced row echelon form corresponding to the matrix  $A - \lambda E$ , especially when  $A$  is a  $n \times n$  matrix with large  $n$ .

We compute

$$|A - \lambda E| = \begin{vmatrix} -\lambda & 0 & 1 \\ 0 & 1 - \lambda & 0 \\ 1 & 0 & -\lambda \end{vmatrix} = -\lambda^3 + \lambda^2 + \lambda - 1,$$

with roots  $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = -1$ . The eigenvectors with eigenvalue  $\lambda = 1$  can be computed:

$$\begin{pmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix};$$

with the basis of the space of solutions, that is, of all eigenvectors with this eigenvalue

$$u_1 = (0, 1, 0), \quad u_2 = (1, 0, 1).$$

Similarly for  $\lambda = -1$  we obtain the third independent eigenvector

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \Rightarrow u_3 = (-1, 0, 1).$$

Under the basis  $u_1, u_2, u_3$  (note that  $u_3$  must be linearly independent of the remaining two because of the previous theorem and  $u_1, u_2$  were obtained as two independent solutions)  $f$  has the diagonal matrix

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

The whole space  $\mathbb{R}^3$  is a direct sum of eigenspaces,  $\mathbb{R}^3 = V_1 \oplus V_2$ , with  $\dim V_1 = 2$ , and  $\dim V_2 = 1$ . This decomposition is uniquely determined and says much about the geometric properties of the mapping  $f$ . The eigenspace  $V_1$  is furthermore a direct sum of one-dimensional eigenspaces, which can be selected in other ways (thus such a decomposition has no further geometrical meaning).

(2) Consider the linear mapping  $f : \mathbb{R}_2[x] \rightarrow \mathbb{R}_2[x]$  defined by polynomial differentiation, that is,  $f(1) = 0$ ,  $f(x) = 1$ ,  $f(x^2) = 2x$ . The mapping  $f$  thus has in the usual basis  $(1, x, x^2)$  the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

The characteristic polynomial is  $|A - \lambda \cdot E| = -\lambda^3$ , thus it has only one eigenvalue,  $\lambda = 0$ . We compute the eigenvectors:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The space of the eigenvectors is thus one-dimensional, generated by the constant polynomial 1.

The striking property of this mapping is that there is no basis for which the matrix would be diagonal. There is the ‘‘chain’’ of vectors mapping the independent generators as follows:  $\frac{1}{2}x^2 \mapsto x \mapsto 1 \mapsto 0$ . This builds a sequence of invariant subspaces without invariant complements.

In case we simply type `A.characteristic_polynomial()`, then Sage returns the characteristic polynomial as a polynomial with respect to the variable  $x$ . The previous list indicates the three eigenvalues of  $A$ , together with their algebraic multiplicity. Of course, one can proceed in a more elementary way, e.g., by the block

```
A=matrix(SR, [[-1, 1, 0], [-1, 3, 0], [2, -2, 2]])
E=identity_matrix(3)
var("c"); ch(c)=det(A-c*E)
ch.factor(); show(ch.roots())
```

Another alternative to compute the eigenvalues of  $A$  has the form

```
A=matrix(SR, [[-1, 1, 0], [-1, 3, 0], [2, -2, 2]])
eigen=A.eigenvalues()
show(eigen)
```

In this case, Sage's output looks like as  $[-\sqrt{3}+1, \sqrt{3}+1, 2]$ , so the command `A.eigenvalues()` computes only the eigenvalues, without their algebraic multiplicities. As for the eigenvectors, add in some of the previous cells the code

```
A.eigenvectors_right()
```

Here Sage prints out the eigenvectors of  $A$ , together with the corresponding eigenvalues and algebraic multiplicities.

**2.D.3.** Find bases of the eigenspaces of the matrix

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

**Solution.** We first compute the determinant  $\det(A - \lambda E)$ :

$$\begin{vmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{vmatrix} = -\lambda \begin{vmatrix} -\lambda & 1 \\ 1 & -\lambda \end{vmatrix} - 1 \begin{vmatrix} 1 & 1 \\ 1 & -\lambda \end{vmatrix} + 1 \begin{vmatrix} 1 & -\lambda \\ 1 & 1 \end{vmatrix}$$

and hence the characteristic polynomial is given by  $\chi_A(\lambda) = -\lambda^3 + 3\lambda + 2$ . With the same method applied in 2.D.1, we see that  $\chi_A(\lambda) = -(\lambda + 1)^2(\lambda - 2)$ , hence  $\lambda_1 = -1$  is a double root, while the other root is given by  $\lambda_2 = 2$ . These are the eigenvalues of  $A$ , with algebraic multiplicities two and one, respectively. Let us describe the associated eigenspaces.

- For  $\lambda_1 = -1$  we get the matrix equation

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Obviously, the reduced row echelon form of the matrix  $A+E$  is given by  $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ , hence we have two free variables,

namely  $x_2 = a \in \mathbb{R}$  and  $x_3 = b \in \mathbb{R}$ . The general solution has the form  $(-a - b, a, b)^T$ , thus the eigenspace  $V_{-1}$  is a 2-dimensional subspace of  $\mathbb{R}^3$  given by

$$V_{-1} = \left\{ \begin{pmatrix} -a-b \\ a \\ b \end{pmatrix} : a, b \in \mathbb{R} \right\} = \text{span}_{\mathbb{R}} \left\{ \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

**2.4.6. Orthogonal mappings.** We consider the special case of the mapping  $f : V \rightarrow W$  between spaces with scalar products, which preserve lengths for all vectors  $u \in V$ .



ORTHOGONAL MAPPINGS

A linear mapping  $f : V \rightarrow W$  between spaces with scalar product is called an *orthogonal mapping*, if for all  $u \in V$

$$\langle f(u), f(u) \rangle = \langle u, u \rangle.$$

The linearity of  $f$  and the symmetry of the scalar product imply that for all pairs of vectors the following equality holds:

$$\langle f(u+v), f(u+v) \rangle = \langle f(u), f(u) \rangle + \langle f(v), f(v) \rangle + 2\langle f(u), f(v) \rangle.$$

Therefore all orthogonal mappings satisfy also the seemingly stronger condition for all vectors  $u, v \in V$ :

$$\langle f(u), f(v) \rangle = \langle u, v \rangle,$$

i.e. the mapping  $f$  leaves the scalar product invariant if and only if it leaves invariant the length of the vectors. (We should have noticed that this is true for all fields of scalars, where  $1 + 1 \neq 0$ , but it does not hold true for  $\mathbb{Z}_2$ .)

In the initial discussion about the geometry in the plane we proved in the Theorem 1.5.10 that a linear mapping  $\mathbb{R}^2 \rightarrow \mathbb{R}^2$  preserves lengths of the vectors if and only if its matrix in the standard basis (which is orthonormal with respect to the standard scalar product) satisfies  $A^T \cdot A = E$ , that is,  $A^{-1} = A^T$ .

In general, orthogonal mappings  $f : V \rightarrow W$  must be always injective, because the condition  $\langle f(u), f(u) \rangle = 0$  implies  $\langle u, u \rangle = 0$  and thus  $u = 0$ . In such a case, the dimension of the range is always at least as large as the dimension of the domain of  $f$ . But then both dimensions are equal and  $f : V \rightarrow \text{Im } f$  is a bijection. If  $\text{Im } f \neq W$ , we extend the orthonormal basis of the image of  $f$  to an orthonormal basis of the range space and the matrix of the mapping then contains a square regular submatrix  $A$  along with zero rows so that it has the required number of rows. Without loss of generality we can assume that  $W = V$ .

Our condition for the matrix of an orthogonal mapping in any orthonormal basis requires that for all vectors  $x$  and  $y$  in the space  $\mathbb{K}^n$ :

$$(A \cdot x)^T \cdot (A \cdot y) = x^T \cdot (A^T \cdot A) \cdot y = x^T \cdot y.$$

Special choice of the standard basis vectors for  $x$  and  $y$  yields directly  $A^T \cdot A = E$ , that is, the same result as for dimension two. Thus we have proved the following theorem:

MATRIX OF ORTHOGONAL MAPPINGS

**Theorem.** Let  $V$  be a real vector space with scalar product and let  $f : V \rightarrow V$  be a linear mapping. Then  $f$  is orthogonal if and only if in some orthonormal basis (and then consequently in all of them) its matrix  $A$  satisfies  $A^T = A^{-1}$ .

These two eigenvectors form a basis of  $V_{-1}$ .

• For  $\lambda_2 = 2$  we obtain the matrix equation

$$\begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

It is easy to verify that the matrix  $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}$  is the re-

duced row echelon form of the matrix  $A - 2E$ . Hence we deduce that the solution has the form  $(a, a, a)^T$ , where  $a = x_3 \in \mathbb{R}$  is the free variable. Thus,  $V_2$  is 1-dimensional,  $V_2 = \text{span}_{\mathbb{R}} \{(1, 1, 1)^T\}$ .  $\square$

**2.D.4.** Present the solution of the task in 2.D.3 in Sage.  $\circ$

*In 2.D.3 we observed that the eigenspace of a multiple eigenvalue, such as the double root  $\lambda_1 = -1$ , can have a dimension greater than one. In fact, for this particular eigenvalue, the algebraic and geometric multiplicity are both equal to 2. However, be aware that, in general, the algebraic and geometric multiplicities of an eigenvalue may not always match. In particular, the algebraic multiplicity of an eigenvalue is always greater than or equal to its geometric multiplicity.*



**2.D.5.** Determine the eigenvalues of the following matrix, together with their algebraic and geometric multiplicities.

$$A = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ 2 & -2 & 2 \end{pmatrix}.$$

**2.D.6.** Compute  $\text{tr}(A)$  and  $\det(A)$  for the matrix

$$A = \begin{pmatrix} -11 & 5 & 4 & 1 \\ -3 & 0 & 1 & 0 \\ -21 & 11 & 8 & 2 \\ -9 & 5 & 3 & 1 \end{pmatrix},$$

based on the formulas

$$\begin{aligned} \det(A) &= \lambda_1 \lambda_2 \lambda_3 \lambda_4, \\ \text{tr}(A) &= \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4, \end{aligned}$$

where  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$  are the roots of the characteristic polynomial  $\chi_A(\lambda)$  of  $A$ . Moreover, do the computation in Sage.

**Solution.** The characteristic polynomial associated to  $A$  is of degree four,  $\chi_A(\lambda) = \lambda^4 + 2\lambda^3 - 2\lambda - 1$ . It has two roots, namely  $\lambda_+ = 1$  with multiplicity one and  $\lambda_- = -1$  with multiplicity three, that is,  $\chi_A(\lambda) = (\lambda - 1)(\lambda + 1)^3$ . It follows that  $\det(A) = \lambda_1 \lambda_2^3 = -1$  and  $\text{tr}(A) = \lambda_1 + 3\lambda_2 = 1 - 3 = -2$ . To perform these computations in Sage, use the following block:

```
A=matrix(SR, [[-11, 5, 4, 1], [-3, 0, 1, 0],
[-21, 11, 8, 2], [-9, 5, 3, 1]])
p(t) = A.characteristic_polynomial(t)
show(p(t).factor())
show(p.roots())
print(A.det()); print(A.trace())
```

**PROOF.** Indeed, if  $f$  preserves lengths, it must have the claimed property in every orthonormal basis. On the other hand, the previous calculations show that this property for the matrix in one such basis ensures length preservation.  $\square$

Square matrices which satisfy the equality  $A^T = A^{-1}$  are called *orthogonal matrices*.

The shape of the coordinate transition matrices between orthonormal bases is a direct corollary of the above theorem. Each such matrix must provide a mapping  $\mathbb{K}^n \rightarrow \mathbb{K}^n$  which preserves lengths and thus satisfies the condition  $S^{-1} = S^T$ . When changing from one orthonormal basis to another one, the matrix of any linear mapping changes according to the relation



$$A' = S^T A S.$$

**2.4.7. Decomposition of an orthogonal mapping.** We take a more detailed look at eigenvectors and eigenvalues of orthogonal mappings on a real vector space  $V$  with scalar product.

Consider a fixed orthogonal mapping  $f : V \rightarrow V$  with the matrix  $A$  in some orthonormal basis. We continue as with the matrix  $D$  of rotation in 2.4.1.

We think first about invariant subspaces of orthogonal mappings and their orthogonal complements. Namely, given any subspace  $W \subset V$  invariant with respect to an orthogonal mapping  $f : V \rightarrow V$ , then for all  $v \in W^\perp$  and  $w \in W$  we immediately see

$$\langle f(v), w \rangle = \langle f(v), f \circ f^{-1}(w) \rangle = \langle v, f^{-1}(w) \rangle = 0$$

since  $f^{-1}(w) \in W$ , too. But this means that also  $f(W^\perp) \subset W^\perp$  and we have proved a simple but very important proposition:

**Proposition.** *The orthogonal complement of a subspace invariant with respect to an orthogonal mapping is also invariant.*

If all eigenvalues of an orthogonal mapping are real, this claim ensures that there always exists a basis of  $V$  composed of eigenvectors. Indeed, the restriction of  $f$  to the orthogonal complement of an invariant subspace is again an orthogonal mapping, therefore we can add one eigenvector to the basis after another, until we obtain the whole decomposition of  $V$ . However, mostly the eigenvalues of orthogonal mappings are not real. We need to deviate into complex vector spaces. We formulate the result right away:




```
eign=A.eigenvalues(); print(eign)
bool(sum(eign)==A.trace())
bool(prod(eign)==A.det())
```

**2.D.7.** Let  $A$  be a  $n \times n$  triangular matrix over  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ , either upper or lower. Prove that the eigenvalues of  $A$  are its diagonal elements.  $\circ$

**2.D.8.** (a) Prove that a  $n \times n$  matrix  $A$  is invertible if and only if 0 is not an eigenvalue of  $A$ .  
 (b) Compute the characteristic polynomial of the  $n \times n$  matrix  $A_n$  given below. Then combine your result with the statement in (a), to determine those  $n$  for which the matrix  $A_n$  is invertible.

$$A_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix}.$$

 As described in 2.4.3 the eigenvectors of a squared matrix corresponding to distinct eigenvalues, are linearly independent. This property enables us to use an “eigenvector basis” to represent the matrix as a diagonal matrix. Matrices (or endomorphisms) with this property are commonly referred to as diagonalizable.

Summarizing, a matrix  $A$  is diagonalizable if it is similar to a diagonal matrix  $D$ , meaning that there exists an invertible matrix  $P$  such that  $A = PDP^{-1}$ . This is equivalent to saying that the algebraic and geometric multiplicities of each eigenvalue of  $A$  are equal.

Diagonalizable matrices are particularly useful because they simplify many matrix computations, such as raising a matrix to a power, or solving systems of linear differential equations. As we will explore in Chapter 3, significant examples of diagonalizable endomorphisms include symmetric matrices. These matrices are important in various fields, including physics, engineering, and computer science.


**2.D.9. Diagonalization.** Consider the matrix  $A$  introduced in the tasks 2.D.1, 2.D.3 and 2.D.5, respectively. Specify the cases where  $A$  is diagonalizable, and for these cases find matrices  $P, D$  such that  $A = PDP^{-1}$ .

**Solution.** (a) Let us consider first the matrix  $A$  given in 2.D.1. There we saw that this matrix admits three eigenvalues, all distinct each other. Hence in this case  $A$  is diagonalizable. In particular, observe that for any eigenvalue of  $A$ , the corresponding algebraic and geometric multiplicity coincide (they are all equal to one). Thus we have  $A = PDP^{-1}$ , with

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_+ & 0 \\ 0 & 0 & \lambda_- \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 2 - \sqrt{3} & 2 + \sqrt{3} \\ 0 & 1 & 1 \\ 1 & -2 & -2 \end{pmatrix},$$

ORTHOGONAL MAPPING DECOMPOSITION

**Theorem.** Let  $f : V \rightarrow V$  be an orthogonal mapping on a real vector space  $V$  with scalar product. Then all the (in general complex) roots of the characteristic polynomial  $f$  have length one. There exists the decomposition of  $V$  into one-dimensional eigenspaces corresponding to the real eigenvalues  $\lambda = \pm 1$  and two-dimensional subspaces  $P_{\lambda, \bar{\lambda}}$  with  $\lambda \in \mathbb{C} \setminus \mathbb{R}$ , where  $f$  acts by the rotation by the angle equal to the argument of the complex number  $\lambda$  in the positive sense. All these subspaces are mutually orthogonal.

 **PROOF.** Without loss of generality we can work with the space  $V = \mathbb{R}^m$  with the standard scalar product. The mapping is thus given by an orthogonal matrix  $A$  which can be equally well seen as the matrix of a (complex) linear mapping on the complex space  $\mathbb{C}^m$  (which just happens to have all of its coefficients real).

There exist exactly  $m$  (complex) roots of the characteristic polynomial of  $A$ , counting their algebraic multiplicities (see the fundamental theorem of algebra, 1.A.6). Furthermore, because the characteristic polynomial of the mapping has only real coefficients, the roots are either real or there are pairs of roots which are complex conjugates  $\lambda$  and  $\bar{\lambda}$ . The associated eigenvectors in  $\mathbb{C}^m$  for such pairs of complex conjugates are actually solutions of two systems of linear homogeneous equations which are also complex conjugate to each other – the corresponding matrices of the systems have real components, except for the eigenvalues  $\lambda$ . Therefore the solutions of this systems are also complex conjugates (check this!).

Next, we exploit the fact that for every invariant subspace its orthogonal complement is also invariant. First we find the eigenspaces  $V_{\pm 1}$  associated with the real eigenvalues, and restrict the mapping to the orthogonal complement of their sum. Without loss of generality we can thus assume that our orthogonal mapping has no real eigenvalues and that  $\dim V = 2n > 0$ .

Now choose an eigenvalue  $\lambda$  and let  $u_\lambda$  be the eigenvector in  $\mathbb{C}^{2n}$  associated to the eigenvalue  $\lambda = \alpha + i\beta$ ,  $\beta \neq 0$ . Analogously to the case of rotation in the plane discussed in paragraph 2.4.1 in terms of the matrix  $D$ , we are interested in the real part of the sum of two one-dimensional (complex) subspaces  $W = \text{span}\{u_\lambda\} \oplus \text{span}\{\bar{u}_\lambda\}$ , where  $\bar{u}_\lambda$  is the eigenvector associated to the conjugated eigenvalue  $\bar{\lambda}$ .

Now we want the intersection of the 2-dimensional complex subspace  $W$  with the real subspace  $\mathbb{R}^{2n} \subset \mathbb{C}^{2n}$ , which is clearly generated (over  $\mathbb{R}$ ) by the vectors  $u_\lambda + \bar{u}_\lambda$  and  $i(u_\lambda - \bar{u}_\lambda)$ . We call this real 2-dimensional subspace  $P_{\lambda, \bar{\lambda}} \subset \mathbb{R}^{2n}$  and notice, this subspace is generated by the basis given by the real and imaginary part of  $u_\lambda$

$$x_\lambda = \text{Re } u_\lambda, \quad -y_\lambda = -\text{Im } u_\lambda.$$

respectively, where  $\lambda_1 = 2, \lambda_{\pm} = 1 \pm \sqrt{3}$  are the eigenvalues of  $A$ . The columns of  $P$  are the eigenvectors of  $A$ , described in 2.D.1. The block below provides a quick verification of the relation  $A = PDP^{-1}$  in Sage:

```
A=matrix(SR, [[-1, 1, 0], [-1, 3, 0],
[2, -2, 2]])
u1=vector([0, 0, 1])
u2=vector([2-sqrt(3), 1, -2])
u3=vector([2+sqrt(3), 1, -2])
P=column_matrix([u1, u2, u3])
D=diagonal_matrix([2, 1+sqrt(3), 1-sqrt(3)])
bool(A==P*D*P.inverse())
```

Note that we can replace any column of the matrix  $P$  with a non-zero scalar multiple of the corresponding eigenvector. As a result, there are infinitely many possible matrices  $P$  that satisfy the diagonalization condition.

(b) Let us now consider the matrix  $A$  given in 2.D.3. There we saw that this matrix has two eigenvalues:  $\lambda_1 = -1$  with algebraic multiplicity 2, and  $\lambda = 2$  with algebraic multiplicity 1. We also proved that the geometric multiplicity of  $\lambda_1$  coincides with its algebraic multiplicity, and similarly for  $\lambda_2$ . Hence in this case the matrix  $A$  is also diagonalizable:  $A = PDP^{-1}$ . To explicitly determine  $D$  and  $P$  we can follow the same method outline in (a). We leave this case as an exercise for practice.

(c) Finally, the matrix  $A$  presented in 2.D.5 is *not* diagonalizable. This is because its unique eigenvalue  $\lambda = 2$  has an algebraic multiplicity 3, but a geometric multiplicity 2. Consequently, it is impossible to construct a complete basis of eigenvectors for this matrix.  $\square$

**2.D.10.** Determine whether the matrices  $A_1$  and  $A_2$  given below are diagonalizable. If diagonalizable, determine the matrices  $D_i, P_i$  such that  $A_i = P_i D_i P_i^{-1}$ :

$$A_1 = \begin{pmatrix} 2 & 0 & 0 \\ 4 & 2 & 2 \\ -2 & 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 2 & 0 & 0 \\ -4 & 2 & 2 \\ -2 & 0 & 1 \end{pmatrix}.$$

Consider a real vector space  $V$  endowed with a scalar product  $\langle \cdot, \cdot \rangle$ , and an orthogonal endomorphism  $f : V \rightarrow V$ . Thus,  $f$  preserves the scalar product, the lengths and the angles of vectors in  $V$ . A prime example of orthogonal endomorphisms are rotations, in  $\mathbb{R}^2$  or  $\mathbb{R}^3$  (cf. 1.E.17).

It is most remarkable, that in matrix terms,  $A \in \text{Mat}_n(\mathbb{R})$  is a matrix of an orthogonal transformation  $f$  (in suitable orthonormal coordinates), if and only if

$$AA^T = A^T A = E.$$

We call them orthogonal matrices and they are particularly easy to invert. This property leads to several interesting results, such as the fact that the determinant of an orthogonal matrix is always  $\pm 1$ .

Because  $A \cdot (u_\lambda + \bar{u}_\lambda) = \lambda u_\lambda + \bar{\lambda} \bar{u}_\lambda$  and similarly with the second basis vector, it is clearly an invariant subspace with respect to multiplication by the matrix  $A$  and we obtain

$$A \cdot x_\lambda = \alpha x_\lambda + \beta y_\lambda, \quad A \cdot (-y_\lambda) = -\alpha y_\lambda + \beta x_\lambda.$$

Consequently,

$$\|A \cdot x_\lambda\|^2 + \|A \cdot y_\lambda\|^2 = (\alpha^2 + \beta^2)(\|x\|^2 + \|y\|^2)$$

and, since our mapping preserves lengths, the absolute value of the eigenvalue  $\lambda$  must equal one. But that means that the restriction of our mapping to  $P_{\lambda, \bar{\lambda}}$  is the rotation by the argument of the eigenvalue  $\lambda$ . Note that the choice of the eigenvalue  $\bar{\lambda}$  instead of  $\lambda$  leads to the same subspace with the same rotation, we would just have expressed it in the basis  $x_\lambda, y_\lambda$ , that is, the same rotation will in these coordinates go by the same angle, but with the opposite sign, as expected.

The proof of the whole theorem is completed by restricting the mapping to the orthogonal complement and finding another 2-dimensional subspace, until we get the required decomposition.  $\square$

We return to the ideas in this proof once again in chapter three, where we study complex extensions of the Euclidean vector spaces, see 3.4.4.

**Remark.** The previous theorem is very powerful in dimension three. Here at least one eigenvalue must be real  $\pm 1$ , since three is odd. But then the associated eigenspace is an axis of the rotation of the three-dimensional space through the angle given by the argument of the other eigenvalues. Try to think how to detect in which direction the space is rotated. Note also that the eigenvalue  $-1$  means an additional reflection through the plane perpendicular to the axis of the rotation.



We shall return to the discussion of such properties of matrices and linear mappings in more details at the end of the next chapter, after illustrating the power of the matrix calculus in several practical applications. We close this section with a general quite widely used definition:

The next task illustrates that orthogonal endomorphisms on  $\mathbb{R}^n$  correspond to orthogonal  $n \times n$  real matrices, and conversely. For more information, see the discussion in 2.4.6.

**2.D.11.** On  $V = \mathbb{R}^n$  endowed with the dot product consider a linear mapping  $\varphi_A : V \rightarrow V$ , defined by  $\varphi_A(u) = Au$ , for some  $A \in \text{Mat}_n(\mathbb{R})$ , for all  $u \in V$ . Prove that  $A$  is orthogonal if and only if  $\varphi_A$  is an orthogonal endomorphism of  $V$ .  $\circ$

**2.D.12.** Determine which of the matrices below is orthogonal:

$$A = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix},$$

$$B = \begin{pmatrix} 2/3 & -2/3 & 1/3 \\ 2/3 & 1/3 & -2/3 \\ 1/3 & 2/3 & 2/3 \end{pmatrix},$$

$$C = \begin{pmatrix} 3/\sqrt{11} & -1/\sqrt{6} & -1/\sqrt{66} \\ 1/\sqrt{11} & 2/\sqrt{6} & -4/\sqrt{66} \\ 1/\sqrt{11} & 1/\sqrt{6} & 7/\sqrt{66} \end{pmatrix}.$$

**Solution.** All the given matrices  $A, B$  and  $C$  are orthogonal. There are many ways to verify this fact. For  $A$  notice that  $A^T = A$  and

$$AA^T = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} = E.$$

On the other hand, later in Chapter 3 we will learn that a square matrix is orthogonal, if and only if its columns (or rows) form an orthonormal basis (see 3.4.4). Given our familiarity with orthonormal bases, we can use this criterion to determine whether the matrix  $B$  is orthogonal. We need to verify for example that the (column) vectors

$$v_1 = \begin{pmatrix} 2/3 \\ 2/3 \\ 1/3 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -2/3 \\ 1/3 \\ 2/3 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 1/3 \\ -2/3 \\ 2/3 \end{pmatrix}$$

form an orthonormal basis of  $\mathbb{R}^3$  with respect to the dot product. We compute  $\|v_1\| = \|v_2\| = \|v_3\| = 1$  and  $v_1 \cdot v_2 = -\frac{4}{9} + \frac{2}{9} + \frac{2}{9} = 0$ ,  $v_1 \cdot v_3 = \frac{2}{9} - \frac{4}{9} + \frac{2}{9} = 0$ ,  $v_2 \cdot v_3 = -\frac{2}{9} - \frac{2}{9} + \frac{4}{9} = 0$ . This proves the claim.

You can also use the command `A.is_unitary()` in Sage. This checks whether a given real matrix  $A$  is orthogonal (or if a complex matrix is unitary, as we will see in Chapter 3). Let us apply this method to verify the orthogonality of the matrix  $C$ :

```
C = matrix(SR, [[3/sqrt(11), -1/sqrt(6), -1/sqrt(66)],
[1/sqrt(11), 2/sqrt(6), -4/sqrt(66)],
[1/sqrt(11), 1/sqrt(6), 7/sqrt(66)]]
C.is_unitary()
```

Sage's output is True.  $\square$

**2.4.8. Definition.** The spectrum of a linear mapping  $f : V \rightarrow V$ , or the spectrum of a square matrix  $A$ , is a sequence of roots of the characteristic polynomial  $f$  or  $A$ , along with their multiplicities, respectively. The algebraic multiplicity of an eigenvalue means the multiplicity of the root of the characteristic polynomial, while the geometric multiplicity of the eigenvalue is the dimension of the associated subspace of eigenvectors.

The spectral diameter of a linear mapping (or matrix) is the greatest of the absolute values of the eigenvalues.

In this terminology, our results about orthogonal mappings can be formulated as follows: the spectrum of an orthogonal mapping is always a subset of the unit circle in the complex plane. Thus only the values  $\pm 1$  may appear in the real part of the spectrum and their algebraic and geometric multiplicities are always the same. Complex values of the spectrum then correspond to rotations in suitable two-dimensional subspaces which are mutually perpendicular.





Proper rotations are orthogonal matrices with determinant 1, and play a crucial role in both applied and pure mathematics. Orthogonal matrices with determinant  $-1$  correspond to reflections through the origin, which are used to model symmetry operations, such as those seen in particle physics or optics. Recall that in low dimensions, eigentheory provides a nuanced geometric perspective on both rotations and reflections (see the remark in 2.4.7, and the tasks 1.E.17, 2.E.57).

In a three-dimensional space, a general proper rotation is characterized by its axis and its angle  $\theta$ . Conventionally, a positive angle corresponds to a counterclockwise rotation when viewed from the tip of the axis, see also the discussion in ???. The rotation matrix in this case has necessarily three eigenvalues, a real one  $\lambda_1 = 1$ , and two unitary complex conjugate eigenvalues.

As explained in 2.4.7, the eigenvector  $u_1$  associated with the eigenvalue 1 represents the rotation axis, which remains invariant under the rotation. The direction of this axis can be determined using the right-hand rule: curl the fingers of your right hand around the axis of rotation, with your fingers pointing in the direction of  $\theta$ ; your thumb will then point in the direction of  $u_1$ . Additionally, the angle  $\theta$  is given by the argument of the two complex conjugate eigenvalues, which represent a 2-dimensional rotation within the plane orthogonal to the rotation axis.

**2.D.13.** Consider the matrix

$$A = \begin{pmatrix} \frac{3}{5} & \frac{16}{25} & \frac{-12}{25} \\ \frac{-16}{25} & \frac{93}{125} & \frac{24}{125} \\ \frac{12}{25} & \frac{24}{125} & \frac{107}{125} \end{pmatrix}.$$

- (a) Show that the linear mapping induced by  $A$  is a rotation.  
 (b) Find its axis, the plane rotation, and the rotation angle.

**Solution.** (a) To prove that the induced linear mapping is a rotation, it is sufficient to show that  $A$  is an orthogonal matrix with determinant 1. In Sage you may apply the method described above, i.e.,

```
A=matrix(SR, [[3/5, 16/25, -12/25],
[ -16/25, 93/125, 24/125],
[12/25, 24/125, 107/125]])
A.is_unitary()
```

Sage's output is True. Adding the command `det(A)`, Sage gives that  $\det(A) = 1$ . Compute the determinant in a formal way, as well. Thus,  $A$  is an orthogonal matrix with determinant 1. Let us also compute the eigenvalues of  $A$  by adding the command `A.eigenvalues()`. Sage returns the list

```
[-4/5*I + 3/5, 4/5*I + 3/5, 1]
```

but for our convenience we set  $\lambda_1 = 1$ ,  $\lambda_2 = \frac{3}{5} + \frac{4}{5}i$  and  $\lambda_3 = \frac{3}{5} - \frac{4}{5}i$ . All three eigenvalues have absolute value one. Combined with the fact that the matrix is orthogonal, this confirms that the matrix represents a rotation.

- (b) In this part we need to find the eigenvectors of  $A$ . Let

us use Sage to specify them. By adding the command `A.eigenvectors_right()` to the previous block, Sage will return the eigenvectors corresponding to the eigenvalues  $\lambda_1, \lambda_2, \lambda_3$ . These are given by  $v_1 = (0, 1, 4/3)^T$ ,  $v_2 = (1, 4i/5, -3i/5)^T$ , and  $v_3 = (1, -4i/5, 3i/5)^T$ , respectively. The axis of the rotation is given by the eigenvector  $v_1$ . The plane of rotation is the real plane  $\mathcal{E}$  in  $\mathbb{R}^3$  which is defined by the intersection of the 2-dimensional complex subspace of  $\mathbb{C}^3$  spanned by the remaining eigenvectors  $v_2, v_3$ , with  $\mathbb{R}^3$ . A direct computation shows that

$$\mathcal{E} = \text{span}_{\mathbb{R}} \{E_1 = (1, 0, 0)^T, E_2 = (0, -4, 3)^T\} \subset \mathbb{R}^3,$$

where the first generator is the real multiple of  $v_2 + v_3$  and the second one is the real multiple of  $i(v_2 - v_3)$ , see also 2.4.7. We are now ready to determine the rotation angle in this plane: It is a rotation by the angle  $\arccos(\frac{3}{5}) \doteq 0, 295\pi$ , which is the argument of the eigenvalue  $\frac{3}{5} + \frac{4}{5}i$  (or minus that number, if we would choose the other eigenvalue).  $\square$

---

*You can find more exercises on eigenvalues, eigenvectors, orthogonal matrices and rotations in Section E. In Chapter 3 we will return to these topics and expand our perspective by exploring their applications in eigenvalues and orthogonal diagonalization.*

---

E. Additional exercises for the whole chapter

Below we provide additional material for most of the concepts discussed in the second chapter. The goal is to gain further computational experience with matrix calculus, become more familiar with basic notions of linear transformations, and hopefully enrich our approach to solving linear problems.

A) MATERIAL ON VECTORS AND MATRICES

**2.E.1. Sage’s vocabulary in linear algebra.** This chapter has provided extensive information on using Sage with vectors and matrices. Before we begin this supplementary section, it is helpful to summarize this knowledge. Below, we list most of the Sage commands we have explored so far, along with instructions for manipulating matrices and linear mappings. We strongly recommend using Sage to aid your understanding of the problems presented below, especially in cases where our solution does not include Sage. This practice will also help expand your Sage vocabulary on linear algebra and its applications, as we will see later in Chapter 3.



Sage vocabulary for matrix calculus, vector spaces and linear mappings

zero $m/n$ matrix	<code>zero_matrix(KK, m, n)</code>	rank	<code>A.rank()</code>
identity $n/n$ matrix	<code>identity_matrix(KK, n)</code>	left kernel	<code>A.left_kernel()</code>
matrix	<code>diagonal_matrix(KK, [a, b, c])</code>	right kernel	<code>A.right_kernel()</code>
random matrix	<code>random_matrix(KK, m, n)</code>	rank of left kernel	<code>A.left_nullity()</code>
matrix multipl.	<code>A * b</code> , or <code>A * B</code> or <code>B * A</code>	rank of right kernel	<code>A.right_nullity()</code>
matrix exponent.	e.g., <code>A^3</code> , or <code>A * *3</code>	character. polyn.	<code>A.characteristic_polynomial()</code>
determinant, trace	<code>A.det()</code> , <code>A.trace()</code>	eigenvalues	<code>A.eigenvalues()</code>
row echelon form	<code>A.echelon_form()</code>	eigenvectors	<code>A.eigenvectors_right()</code>
RREF	<code>A.rref()</code>	conjugate transpose	<code>A.conjugate_transpose()</code>
pivots columns	<code>A.pivots()</code>	orthogonal	<code>A.is_unitary()</code>
inverse	<code>A.inverse()</code>	unitary	<code>A.is_unitary()</code>
minors of order $k$	<code>A.minor(k)</code>	symmetric	<code>A.is_symmetric()</code>
adjoint	<code>A.adjugate()</code>	rescale rows	<code>A.rescale_row()</code>
transpose	<code>A.transpose()</code>	add multiples of rows	<code>A.add_multiple_of_row()</code>
conjugate	<code>A.conjugate()</code>	interchange rows	<code>A.swap_rows()</code>
accessing the entry $a_{1,2}$	<code>A[0, 1]</code> or <code>A[0][1]</code>	find the size of $A$	<code>A.rows()</code> , <code>A.ncols()</code>
accessing $k$ th column	<code>A[:, k - 1]</code> , e.g. 3rd column <code>A[:, 2]</code>	introduce a vector $u$	<code>u = vector(KK, [u1, ..., un])</code>
accessing $n$ th row	<code>A[n - 1, :]</code> , e.g. 2nd row <code>A[1, :]</code>	size of $u$	<code>u.degree()</code>
submatrices of $B$	<code>B[0 : 2, 1 : 3]</code> , <code>B[0 : 2, 1 : 3]</code>	block matrix $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$	<code>A.stack(B)</code>
e.g., $B$ is a $4/5$ matrix	<code>B[:, 1 : 2]</code> , <code>B[:, 1 : 3]</code> <code>B[1 : 2, :]</code> , <code>B[1 : 3, :]</code>	augmented matrix	<code>A.augment(b)</code>
		a solution of $Ax = b$	<code>A\b</code>

As we learned in the main part, many of the commands listed above come with various options, which we have not included here for simplicity. Above,  $\mathbb{K}$  represents a numerical system  $\mathbb{K}$ , e.g.  $\mathbb{Z}$  for the integers  $\mathbb{Z}$ , etc.

**2.E.2.** Consider the finite field  $\mathbb{Z}_7$  and the matrices  $A = \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix}$  and  $B = \begin{pmatrix} 6 & 1 \\ 0 & 2 \end{pmatrix}$  with elements in  $\mathbb{Z}_7$ . Describe the multiplication  $AB$  and next confirm your result using Sage. How many arithmetic operations we need to compute the product  $AB$  over  $\mathbb{Z}_7$  and why this differs from the result presented in 1.E.11?

**Solution.** The finite field  $\mathbb{Z}_7$  consists of the integers  $\{0, 1, 2, 3, 4, 5, 6\}$  with addition and multiplication defined modulo 7. Thus, to multiply the matrices  $A, B$  over  $\mathbb{Z}_7$  we follow the standard procedure for matrix multiplication but reduce each entry modulo 7. Let us denote by  $C = AB = (c_{ij})$  the product of  $A, B$ . Then, the entries  $c_{ij}$  are given by  $c_{ij} = \sum_{k=1}^2 a_{ik}b_{kj} \pmod{7}$ . We compute

$$c_{11} = (2 \cdot 6 + 3 \cdot 0) \pmod{7} = 12 \pmod{7} = 5, \quad c_{12} = (2 \cdot 1 + 3 \cdot 2) \pmod{7} = 8 \pmod{7} = 1,$$

$$c_{21} = (4 \cdot 6 + 5 \cdot 0) \pmod{7} = 24 \pmod{7} = 3, \quad c_{22} = (4 \cdot 1 + 5 \cdot 2) \pmod{7} = 14 \pmod{7} = 0.$$

Thus

$$C = AB = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 3 & 0 \end{pmatrix}.$$

A verification in Sage follows the standard procedure. However, we first need to introduce the finite field  $\mathbb{Z}_7$  and define the matrices  $A$  and  $B$  over this field.

$Z7=GF(7)$ 
 $A=Matrix(Z7, [[2, 3], [4, 5]]); B=Matrix(Z7, [[6, 1], [0, 2]]); show(A*B)$ 

Finally, according to 1.E.11 the multiplication  $AB$  of two  $2 \times 2$  real matrices requires 12 arithmetic operations: For each matrix element, we perform two multiplications and one addition. Over  $\mathbb{Z}_7$  for each matrix element we need one more operation, the modulo operation. Thus in total we will have  $12 + 4 = 16$  operations.  $\square$

*Elementary matrices are fundamental tools in linear algebra, particularly useful for understanding and performing row operations on matrices. An elementary matrix is derived from the identity matrix by performing a single elementary row operation, see 2.1.8. These operations are essential for matrix manipulations, solving systems of linear equations, and finding matrix inverses. Let us describe related tasks.*

2.E.3. Consider the matrix  $A = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & -1 \\ 2 & 4 & 2 \end{pmatrix}$ .

- Determine the elementary matrix representing the row operation  $R_2 \rightarrow R_2 - 3R_1$ . Next, confirm that the inverse of this elementary matrix is the elementary matrix of the row operation  $R_2 \rightarrow R_2 + 3R_1$ .
- Determine the elementary matrix representing the row operation  $R_2 \rightarrow R_2 - \frac{3}{2}R_3$ . Next, confirm that the inverse of this elementary matrix is the elementary matrix of the row operation  $R_2 \rightarrow R_2 + \frac{3}{2}R_3$ .
- Determine the elementary matrix representing the row operation  $R_3 \rightarrow \frac{1}{2}R_3$ . Next, confirm that the inverse of this elementary matrix is the elementary matrix of the row operation  $R_3 \rightarrow 2R_3$ .  $\circ$

2.E.4. Consider the matrix  $A = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 0 & 1 \\ 4 & 2 & 1 \end{pmatrix}$ .

- Show that  $A$  is invertible.
- Apply elementary row operations to determine the inverse of  $A$ .
- Express  $A$  as a product of inverses of elementary matrices.

**Solution.** (a) Using the Sarrus rule (see 2.2.1) it is straightforward to verify that the determinant of  $A$  equals  $5 \neq 0$ . Therefore, the matrix is invertible.

(b) Consider the augmented matrix  $(A \mid E)$  (where  $E$  is the  $3 \times 3$  identity). The goal is to use elementary row operations to transform  $(A \mid E)$  into  $(E \mid A^{-1})$ . We compute:

$$\begin{aligned} (A \mid E) &= \left( \begin{array}{ccc|ccc} 2 & 1 & 3 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 4 & 2 & 1 & 0 & 0 & 1 \end{array} \right) \xrightarrow{R_1 \rightarrow \frac{1}{2}R_1} \left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 4 & 2 & 1 & 0 & 0 & 1 \end{array} \right) \xrightarrow{R_2 \rightarrow R_2 - R_1} \left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 & 0 \\ 4 & 2 & 1 & 0 & 0 & 1 \end{array} \right) \\ &\xrightarrow{R_3 \rightarrow R_3 - 4R_1} \left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 & 0 \\ 0 & 0 & -5 & -2 & 0 & 1 \end{array} \right) \xrightarrow{R_2 \rightarrow -2R_2} \left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 1 & 1 & -2 & 0 \\ 0 & 0 & -5 & -2 & 0 & 1 \end{array} \right) \xrightarrow{R_3 \rightarrow -\frac{1}{5}R_3} \\ &\left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 1 & 1 & -2 & 0 \\ 0 & 0 & 1 & \frac{2}{5} & 0 & -\frac{1}{5} \end{array} \right) \xrightarrow{R_2 \rightarrow R_2 - R_3} \left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & \frac{3}{5} & -2 & \frac{1}{5} \\ 0 & 0 & 1 & \frac{2}{5} & 0 & -\frac{1}{5} \end{array} \right) \xrightarrow{R_1 \rightarrow R_1 - \frac{3}{2}R_3} \\ &\left( \begin{array}{ccc|ccc} 1 & \frac{1}{2} & 0 & -\frac{1}{10} & 0 & \frac{3}{10} \\ 0 & 1 & 0 & \frac{3}{5} & -2 & \frac{1}{5} \\ 0 & 0 & 1 & \frac{2}{5} & 0 & -\frac{1}{5} \end{array} \right) \xrightarrow{R_1 \rightarrow R_1 - \frac{1}{2}R_2} \left( \begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{2}{5} & 1 & \frac{1}{5} \\ 0 & 1 & 0 & \frac{3}{5} & -2 & \frac{1}{5} \\ 0 & 0 & 1 & \frac{2}{5} & 0 & -\frac{1}{5} \end{array} \right). \end{aligned}$$

Thus  $A^{-1} = \begin{pmatrix} -2/5 & 1 & 1/5 \\ 3/5 & -2 & 1/5 \\ 2/5 & 0 & -1/5 \end{pmatrix}$ . To confirm this expression, you can either use Sage, or verify the relation  $AA^{-1} =$

$E = A^{-1}A$  by hand. Let us present the verification in Sage:

```
A=Matrix(SR, [[2, 1, 3], [1, 0, 1], [4, 2, 1]]); print(det(A))
A_inv=A.inverse(); show(A_inv)
```

(c) In part (b) we obtained  $A^{-1}$  by applying eight elementary row operations. These operations correspond to eight elementary matrices  $E_1, \dots, E_8$ , such that  $E_8 E_7 E_6 E_5 E_4 E_3 E_2 E_1 A = E$ . Thus,  $A = (E_8 E_7 E_6 E_5 E_4 E_3 E_2 E_1)^{-1} = E_1^{-1} E_2^{-1} E_3^{-1} E_4^{-1} E_5^{-1} E_6^{-1} E_7^{-1} E_8^{-1}$ . Let us specify  $E_1, \dots, E_8$  and their inverses.

$$\text{row operation } R_1 \rightarrow \frac{1}{2}R_1 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_1 \rightarrow \frac{1}{2}R_1} \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} =: E_1 \implies E_1^{-1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\text{row operation } R_2 \rightarrow R_2 - R_1 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - R_1} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} =: E_2 \implies E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\text{row operation } R_3 \rightarrow R_3 - 4R_1 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_3 \rightarrow R_3 - 4R_1} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & 0 & 1 \end{pmatrix} =: E_3 \implies E_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix},$$

$$\text{row operation } R_2 \rightarrow -2R_2 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow -2R_2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix} =: E_4 \implies E_4^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\text{row operation } R_3 \rightarrow -\frac{1}{5}R_3 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_3 \rightarrow -\frac{1}{5}R_3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1/5 \end{pmatrix} =: E_5 \implies E_5^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -5 \end{pmatrix},$$

$$\text{row operation } R_2 \rightarrow R_2 - R_3 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - R_3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} =: E_6 \implies E_6^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\text{row operation } R_1 \rightarrow R_1 - \frac{3}{2}R_3 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_1 \rightarrow R_1 - \frac{3}{2}R_3} \begin{pmatrix} 1 & 0 & -\frac{3}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} =: E_7 \implies E_7^{-1} = \begin{pmatrix} 1 & 0 & \frac{3}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\text{row operation } R_1 \rightarrow R_1 - \frac{1}{2}R_2 : E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_1 \rightarrow R_1 - \frac{1}{2}R_2} \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} =: E_8 \implies E_8^{-1} = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is now a straightforward (though lengthy) computation to verify that  $A = E_1^{-1} E_2^{-1} E_3^{-1} E_4^{-1} E_5^{-1} E_6^{-1} E_7^{-1} E_8^{-1}$ . To speed up this process, we can program Sage to perform the calculation for us, by adding the following cell to the previous code block:

```
E1inv=Matrix(SR, [[2, 0, 0], [0, 1, 0], [0, 0, 1]])
E2inv=Matrix(SR, [[1, 0, 0], [1, 1, 0], [0, 0, 1]])
E3inv=Matrix(SR, [[1, 0, 0], [0, 1, 0], [4, 0, 1]])
E4inv=Matrix(SR, [[1, 0, 0], [0, -1/2, 0], [0, 0, 1]])
E5inv=Matrix(SR, [[1, 0, 0], [0, 1, 0], [0, 0, -5]])
E6inv=Matrix(SR, [[1, 0, 0], [0, 1, 1], [0, 0, 1]])
E7inv=Matrix(SR, [[1, 0, 3/2], [0, 1, 0], [0, 0, 1]])
E8inv=Matrix(SR, [[1, 1/2, 0], [0, 1, 0], [0, 0, 1]])
bool(A==E1inv*E2inv*E3inv*E4inv*E5inv*E6inv*E7inv*E8inv)
```

Execute this program yourselves to verify that Sage's output is True. □

*Sage provides built-in functions to explore elementary matrices, which we can list as follows:*

- *The matrix which multiplies the  $k$ th row by  $c$ :*  
`elementary_matrix(R, n, row1 = k, scale = c)`
- *The matrix which multiplies the  $k$ th row by  $c$  and adds it to the  $m$ th row:*  
`elementary_matrix(R, n, row1 = k, row2 = m, scale = c)`
- *The matrix which swaps the  $k$ th and  $m$ th rows:*  
`elementary_matrix(R, n, row1 = k, row2 = m)`

*In each case,  $R$  is the base ring, and is optional, while  $n$  denotes the size of the square matrix.*

**2.E.5.** Use Sage to describe the elementary matrices  $E_1, \dots, E_8$  obtained in 2.E.4, using the commands mentioned above. Compute their inverses, as well. ○

**2.E.6.** Solve the system of linear equations given by the extended matrix  $B = (A \mid b) = \left( \begin{array}{cccc|c} 3 & 3 & 2 & 1 & 3 \\ 2 & 1 & 1 & 0 & 4 \\ 0 & 5 & -4 & 3 & 1 \\ 5 & 3 & 3 & -3 & 5 \end{array} \right)$ . ○

**2.E.7.** Provide a formal solution of the linear system given below. Next use Sage to verify your solution:

$$\begin{cases} x_1 + x_2 + x_3 + x_4 - 2x_5 = 3, \\ 2x_2 + 2x_3 + 2x_4 - 4x_5 = 5, \\ -x_1 - x_2 - x_3 + x_4 + 2x_5 = 0, \\ -2x_1 + 3x_2 + 3x_3 - 6x_5 = 2. \end{cases}$$

**Solution.** The extended matrix of the system is

$$B = (A \mid b) = \left( \begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 2 & 2 & 2 & -4 & 5 \\ -1 & -1 & -1 & 1 & 2 & 0 \\ -2 & 3 & 3 & 0 & -6 & 2 \end{array} \right).$$

Adding the first row to the third, adding its 2-multiple to the fourth, and adding the  $(-5/2)$ -multiple of the second to the fourth we obtain

$$B \sim \left( \begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 2 & 2 & 2 & -4 & 5 \\ 0 & 0 & 0 & 2 & 0 & 3 \\ 0 & 5 & 5 & 2 & -10 & 8 \end{array} \right) \sim \left( \begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 2 & 2 & 2 & -4 & 5 \\ 0 & 0 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & -3 & 0 & -9/2 \end{array} \right).$$

The last row is clearly a multiple of the previous one, and thus we can omit it. The pivots are located in the first, second and fourth, while  $x_3$  and  $x_5$  are free variables, which we substitute by the real parameters  $t$  and  $s$ . Thus we consider the system

$$\begin{aligned} x_1 + x_2 + t + x_4 - 2s &= 3, \\ 2x_2 + 2t + 2x_4 - 4s &= 5, \\ 2x_4 &= 3. \end{aligned}$$

We see that  $x_4 = 3/2$ , and hence the second equation gives  $x_2 = 1 - t + 2s$ . From the first equation we now obtain  $x_1 = 1/2$ . This gives the expression  $(x_1, x_2, x_3, x_4, x_5) = (1/2, 1 - t + 2s, t, 3/2, s)$  with  $t, s \in \mathbb{R}$ . Note that setting  $t_0 = -t + 2s$  and  $t_1 = t - s$ , the solution takes the forms  $(1/2, 1 + t_0, t_0 + 2t_1, 3/2, t_0 + t_1)$ , with  $t_0, t_1 \in \mathbb{R}$ .

As an alternative method, one could transform the extended matrix  $B$  to reduced row echelon form. In this procedure, we can omit the fourth equation since it is a combination of the first three. By sequentially multiplying the second and third rows by  $1/2$ , subtracting the third row from the second and first rows, and finally subtracting the second row from the first, we obtain the following:

$$B \sim \left( \begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 2 & 2 & 2 & -4 & 5 \\ 0 & 0 & 0 & 2 & 0 & 3 \end{array} \right) \sim \left( \begin{array}{ccccc|c} 1 & 1 & 1 & 1 & -2 & 3 \\ 0 & 1 & 1 & 1 & -2 & 5/2 \\ 0 & 0 & 0 & 1 & 0 & 3/2 \end{array} \right) \sim \left( \begin{array}{ccccc|c} 1 & 1 & 1 & 0 & -2 & 3/2 \\ 0 & 1 & 1 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 & 0 & 3/2 \end{array} \right) \sim \left( \begin{array}{ccccc|c} 1 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 1 & 1 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 & 0 & 3/2 \end{array} \right).$$

This confirms that  $x_3, x_5$  are free variables, i.e.,  $x_3 = t, x_5 = s$  with  $t, s \in \mathbb{R}$ . Moreover, a short computation yields the solution presented above. These computations, particularly the (unique) reduced row echelon form of the extended matrix  $B$  and the given solution, can be obtained in Sage by applying the method proposed in 2.A.13. In particular, the commands `B.rref()` and `B.pivots()` appearing in the block below, return the reduced row echelon form of  $B$  and the pivot columns, respectively.

```
A=matrix([[1, 1, 1, 1, -2], [0, 2, 2, 2, -4],
[-1, -1, -1, 1, 2], [-2, 3, 3, 0, -6]])
b=vector([3, 5, 0, 2]); B=A.augment(b)
Br=B.rref(); show(Br); B.pivots()
x=A.solve_right(b)
k=A.right_kernel().matrix()
nrows = k.nrows()
```

```

if nrows > 0:
    t = vector ( var("t", n= nrows ))
    show (x + t*k)
else:
    show (x)
    
```

Sage's output has the form  $(\frac{1}{2}, t_0 + 1, t_0 + 2t_1, \frac{3}{2}, t_0 + t_1)$ , with  $t_0, t_1 \in \mathbb{R}$ , and this verifies our formal computations.  $\square$

**2.E.8.** Find all the solutions of the following homogeneous system of four linear equations and five unknowns  $x, y, z, u, v$ :

$$\{x + y = 2z + v, \quad z + 4u + v = 0, \quad -3u = 0, \quad z = -v\}.$$

Moreover, specify a basis of the vector space of solutions.

**Solution.** Let us first rewrite the system into the matrix form  $A\mathbf{x} = 0$ , where here  $\mathbf{x} = (x, y, z, u, v)^T \in \mathbb{R}^5$ , and as usual the first column of  $A$  consists of the coefficients of  $x$ , the second column of the coefficients of  $y$ , and so on. This gives

$$A = \begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Such homogeneous systems with less (linear) equations than unknowns are called *underdetermined* and admit an infinite number of solutions. Indeed, let us compute the reduced row echelon form of  $A$ . For this, let us denote by  $R_1, R_2, R_3, R_4$  the rows at any step of the procedure. We see that

$$\begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 + \frac{4}{3} \cdot R_3} \begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{R_4 \rightarrow R_4 - R_2} \begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\xrightarrow{R_3 \rightarrow -\frac{1}{3} \cdot R_3} \begin{pmatrix} 1 & 1 & -2 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \rightarrow R_1 + 2 \cdot R_2} \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The final matrix is in reduced row echelon form. Recall that in Sage we can achieve this form in two ways: by applying the command `A.rref()`, or by manually performing the necessary elementary row operations. Let us demonstrate both methods using matrices over  $\mathbb{Q}$  instead of  $\mathbb{R}$ , for convenience, as this not affect the result.

```

A=matrix(QQ, [[1, 1, -2, 0, -1], [0, 0, 1, 4, 1],
[0, 0, 0, -3, 0],[0, 0, 1, 0, 1]])
show(A.rref())
    
```

The output is the same as the one obtained by executing the following block:

```

A=matrix(QQ, [[1, 1, -2, 0, -1], [0, 0, 1, 4, 1],
[0, 0, 0, -3, 0],[0, 0, 1, 0, 1]])
A.add_multiple_of_row(1, 2, 4/3);show(A)
A.add_multiple_of_row(3, 1, -1);show(A)
A.rescale_row(2, -1/3);show(A)
A.add_multiple_of_row(0, 1, 2);show(A)
    
```

where recall that elementary row operations should be applied successively.

Running any of these blocks verifies that our computations are accurate. Therefore, we can now pose the system  $\{x + y + v = 0, z + v = 0, u = 0\}$ , where the variable  $y, v$  are free (they correspond to the 2nd and 5th column of the REEF, since these columns do not contain a pivot). Hence, let us set  $y = t$ , and  $v = s$  with  $t, s \in \mathbb{R}$  and derive the solution by back substitution:

$$(x, y, z, u, v) = (-t - s, t, -s, 0, s), \quad t, s \in \mathbb{R}.$$

Obviously, we can rewrite the solution as

$$\begin{pmatrix} x \\ y \\ z \\ u \\ v \end{pmatrix} = t \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} -1 \\ 0 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \quad t, s \in \mathbb{R}.$$

It easy to prove that the vectors in the right-hand side are linearly independent, and thus they form a basis of the solution space. In other words, the kernel of the matrix  $A$  is a 2-dimensional subspace of  $\mathbb{R}^5$ .  $\square$

**2.E.9.** Determine all solutions of the following linear system:

$$\begin{cases} x_2 + x_4 = 1, \\ 3x_1 - 2x_2 - 3x_3 + 4x_4 = -2, \\ x_1 + x_2 - x_3 + x_4 = 2, \\ x_1 - x_3 = 1. \end{cases}$$

**2.E.10.** Solve the following linear system:

$$\begin{cases} 3x - 5y + 2u + 4z = 2, \\ 5x + 7y - 4u - 6z = 3, \\ 7x - 4y + u + 3z = 4, \\ x + 6y - 2u - 5z = 2. \end{cases}$$

**2.E.11.** Consider the following linear system:

$$\begin{cases} x_1 + x_2 - x_3 = 1, \\ x_1 + 2x_2 + \kappa x_3 = 2, \\ 2x_1 + \kappa x_2 + 2x_3 = 3. \end{cases}$$

Determine the values of the real parameter  $\kappa$  such that

- (a) the system has a unique solution;
- (b) the system has infinitely many solutions;
- (c) the system has no solution.

For the cases (a) and (b) determine the corresponding solutions.

**Solution.** The extended matrix of the system has the form

$$B = (A \mid b) = \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 1 & 2 & \kappa & 2 \\ 2 & \kappa & 2 & 3 \end{array} \right).$$

By applying elementary row operations we obtain that

$$B = \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 1 & 2 & \kappa & 2 \\ 2 & \kappa & 2 & 3 \end{array} \right) \begin{array}{l} R_2 \rightarrow R_2 - R_1 \\ R_3 \rightarrow R_3 - 2R_1 \end{array} \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \kappa + 1 & 1 \\ 0 & \kappa - 2 & 4 & 1 \end{array} \right) \begin{array}{l} R_3 \rightarrow R_3 - (\kappa - 2)R_2 \end{array} \\ \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \kappa + 1 & 1 \\ 0 & 0 & 4 - (\kappa + 1)(\kappa - 2) & 3 - \kappa \end{array} \right).$$

The above matrix is in row echelon form and we can rewrite it as

$$B \sim \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \kappa + 1 & 1 \\ 0 & 0 & -\kappa^2 + \kappa + 6 & 3 - \kappa \end{array} \right) = \left( \begin{array}{ccc|c} 1 & 1 & -1 & 1 \\ 0 & 1 & \kappa + 1 & 1 \\ 0 & 0 & (\kappa - 3)(\kappa + 2) & 3 - \kappa \end{array} \right). \quad (*)$$

Hence we deduce that:

(a) The system has a unique solution when  $\kappa \neq 3$  and  $\kappa \neq -2$ . Indeed, in this case we obtain

$$\begin{cases} x_1 + x_2 - x_3 = 1, \\ x_2 + (\kappa + 1)x_3 = 1, \\ (2 + \kappa)(3 - \kappa)x_3 = 3 - \kappa. \end{cases}$$

Therefore, by back substitution we obtain the solution  $(x_1, x_2, x_3)^T = (1, \frac{1}{2+\kappa}, \frac{1}{2+\kappa})^T$ .

(b) The system has infinite many solutions when  $\kappa = 3$ , since then the last row of the row echelon form presented in (\*) has only zeros, the pivots belong to the first and second column, and  $x_3$  is a free variable. Hence in this case we obtain the system

$$\begin{cases} x_1 + x_2 - x_3 = 1, \\ x_2 + 4x_3 = 1, \end{cases}$$

with  $x_3 = t \in \mathbb{R}$ , having infinite many solutions given by  $(x_1, x_2, x_3)^T = (5t, 1 - 4t, t)^T$ , ( $t \in \mathbb{R}$ ). Note that we may express these solutions as  $(x_1, x_2, x_3)^T = t(5, -4, 1)^T + (0, 1, 0)^T$ . Thus, when  $\kappa = 3$  the space of all solutions is given by summing all solutions of the corresponding homogeneous system, given by  $t(5, -4, 1)^T$ , with a particular solution of the original system, given by  $(0, 1, 0)^T$ .



(c) The system is inconsistent when  $\kappa = -2$ . Indeed, for this value of  $\kappa$  the third row of the matrix in the right-hand side of (\*) gives us  $0x_3 = 5$ , which is impossible.  $\square$

2.E.12. Solve the following parametric system of linear equations in terms of the parameter  $\mu \in \mathbb{R}$ :

$$\begin{cases} \mu x_1 + 4x_2 + 2x_3 = 0, \\ 2x_1 + 3x_2 - x_3 = 0. \end{cases}$$

2.E.13. Determine the number of solutions of the following parametric linear system, in terms of the parameter  $a \in \mathbb{R}$ :

$$\begin{pmatrix} 4 & 1 & 4 & a \\ 2 & 3 & 6 & 8 \\ 3 & 2 & 5 & 4 \\ 6 & -1 & 2 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 3 \\ -3 \end{pmatrix}.$$

2.E.14. Consider the following parametric system of linear equations:

$$\begin{cases} x_1 - a x_2 - 2x_3 = b, \\ x_1 + (1-a)x_2 = b-3, \\ x_1 + (1-a)x_2 + a x_3 = 2b-1. \end{cases}$$

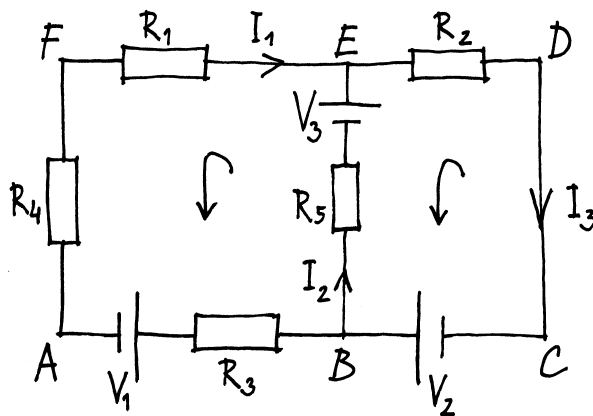
Find the values of the parameters  $a, b \in \mathbb{R}$ , such that:

- (a) the system has exactly one solution;
- (b) the system has no solution;
- (c) the system has infinitely many solutions.



Let us now explore an important application of linear systems in the analysis of electric circuits. This description relies on Ohm's law, the Kirchhoff's voltage and the current laws, providing an opportunity to revisit some concepts from high school physics. Readers not interested in these applications can skip this task.

2.E.15. **Kirchhoff's Circuit Laws.** Consider an electric circuit as in the figure and write down the values of the currents there if you know the values  $V_1 = 20$ ,  $V_2 = 120$ ,  $V_3 = 50$ ,  $R_1 = 10$ ,  $R_2 = 30$ ,  $R_3 = 4$ ,  $R_4 = 5$ ,  $R_5 = 10$ . Notice that the quantities  $I_i$  denote the electric currents, while  $R_j$  are resistances, and  $V_k$  are voltages.



**Solution.** There are two closed loops, namely  $ABEF$  and  $EBCD$  and two branching vertices  $B$  and  $E$  of degree no less than 3. On every segment of the circuit, bounded by branching points, the electric current is constant. Let us set this  $I_1$  for the segment  $EFAB$ ,  $I_2$  for  $EB$ , and  $I_3$  for  $BCDE$ .



Applying Kirchhoff's current law to branching points  $B$  and  $E$  we obtain:  $I_1 + I_2 = I_3$  and  $I_3 - I_1 = I_2$ , which are, of course the same equations. In case there are many branching vertices, we write all Kirchhoff's Current Law equations to the system, having at least one of those equations redundant. Choose the counter clockwise orientations of the loops  $ABEF$  and  $EBCD$ . Applying Kirchhoff Voltage Law and Ohm's Law to the loop  $ABEF$  we obtain the equation:

$$V_1 + I_1 R_3 - I_2 R_5 + V_3 + I_1 R_1 + I_1 R_4 = 0.$$

Similarly, the loop  $EBCD$  implies  $-V_2 + I_3R_2 - V_3 + R_5I_2 = 0$ . A combination of all equations yields the following system of linear equations:

$$\begin{cases} I_1 + I_2 - I_3 = 0, \\ (R_3 + R_1 + R_4)I_1 - R_5I_2 + I_3 = -V_1 - V_3, \\ R_5I_2 + R_2I_3 = V_2 + V_3. \end{cases}$$

Using the prescribed values, this finally reduce to

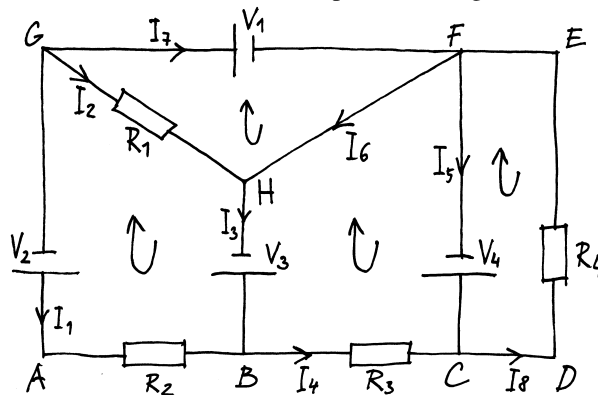
$$\begin{cases} I_1 + I_2 - I_3 = 0, \\ 19I_1 - 10I_2 + I_3 = -70, \\ 10I_2 + 30I_3 = 170. \end{cases}$$

It is easy to see that this has a unique solution, given by  $I_1 = -\frac{80}{53} \approx -1.509$ ,  $I_2 = \frac{219}{53} \approx 4.132$ ,  $I_3 = \frac{139}{53} \approx 2.623$ .  $\square$

**2.E.16. The general case.** In general, the method for electrical circuit analysis can be formulated along the following steps:

- ( $\alpha$ ) Identify all *branching vertices* of the circuit, i.e., vertices of degree no less than 3;
- ( $\beta$ ) Identify all *closed loops* of the circuit;
- ( $\gamma$ ) Introduce variables  $I_k$ , denoting the *oriented currents* on each segment of the circuit between two branching vertices;
- ( $\delta$ ) Write down the *Kirchhoff's current conservation law* for each branching vertex. The total incoming current equals the total outgoing current;
- ( $\varepsilon$ ) Choose an orientation on every closed loop of the circuit and write down the *Kirchhoff's voltage conservation law*, according to the chosen orientation. Here, in case you find an electric charge of voltage  $V_j$  and you go from the short bar to the long bar, then the contribution of this charge is  $V_j$ . It should be  $-V_j$  in case you go from the long bar to the short one. Notice also that if you go in the positive direction of a current  $I$  and find a resistor with resistance  $R_j$ , then the contribution is  $-R_jI$ , and it is  $R_jI$  if the orientation of the loop is opposite to the direction of the current  $I$ . The total voltage change along each closed loop must be zero.
- ( $\zeta$ ) Compose the system collecting all equations, representing the Kirchhoff's current and voltage laws and solve it with respect to the variables, representing the currents. Notice that some equations may be redundant, however, the solution should be unique!

To illustrate this general approach, consider the circuit example in the diagram below.



**Solution.** Let us apply the steps posed above:

- ( $\alpha$ ) The set of branching vertices is  $\{B, C, F, G, H\}$ .
- ( $\beta$ ) The set of closed loops is  $\{ABHG, FHBC, GHF, CDEF\}$ .
- ( $\gamma$ ) Let  $I_1$  be the current on the segment  $GAB$ ,  $I_2$  on the segment  $GH$ ,  $I_3$  on the segment  $HB$ ,  $I_4$  on the segment  $BC$ ,  $I_5$  on the segment  $FC$ ,  $I_6$  on the segment  $FH$ ,  $I_7$  on  $GF$ , and  $I_8$  on  $CDEF$ .
- ( $\delta$ ) Next we write the Kirchhoff's current conservation laws for the branching vertices:

$$B : I_1 + I_3 = I_4, \quad C : I_4 + I_5 = I_8, \quad F : I_8 = I_5 + I_6 - I_7, \quad G : -I_7 = I_1 + I_2, \quad H : I_2 + I_6 = I_3.$$

Notice that the second and third equations give  $I_8 - I_5 = I_4$  and  $I_8 - I_5 = I_6 - I_7$ , respectively. Hence we can replace any of them by the equation  $I_4 = I_6 - I_7$ .

- ( $\varepsilon$ ) Let us now write Kirchhoff's voltage conservation for each of the closed loops traversed counter-clockwise:

$$\begin{aligned} \text{loop } ABHG : -R_1I_2 + V_3 + R_2I_1 - V_2 &= 0, & \text{loop } GHF : R_1I_2 - V_1 &= 0, \\ \text{loop } FHBC : V_4 + R_3I_4 - V_3 &= 0, & \text{loop } CDEF : R_4I_8 - V_4 &= 0. \end{aligned}$$

- ( $\zeta$ ) To pass to the final step we need to use some specific values for  $R_i, V_i$  for  $i = 1, \dots, 4$ . Setting  $R_1 = 4, R_2 = 7, R_3 = 9, R_4 = 12, V_1 = 10, V_2 = 20, V_3 = 60$  and  $V_4 = 120$ , we get a unique solution (although the system is overdetermined!)

$$I_1 = -\frac{30}{7}, I_2 = \frac{5}{2}, I_3 = -\frac{50}{21}, I_4 = -\frac{20}{3}, I_5 = \frac{50}{3}, I_6 = -\frac{205}{42}, I_7 = \frac{25}{14}, I_8 = 10.$$

Below we present the solution in Sage, where first we should introduce the unknowns  $I_1, \dots, I_8$  as symbolic variables.

```
var("I1, I2, I3, I4, I5, I6, I7, I8")
eq1=I1+I3-I4; eq2=I4+I5-I8
eq3=I5+I6-I7-I8; eq4=I1+I2+I7
eq5=I2-I3+I6; eq6=7*I1-4*I2+40
eq7=9*I4+60; eq8=4*I2-10
eq9=12*I8-120
show(solve([eq1, eq2, eq3, eq4, eq5, eq6, eq7, eq8, eq9],
[I1, I2, I3, I4, I5, I6, I7, I8]))
```

□

### B) MATERIAL ON DETERMINANTS

**2.E.17.** Factor the following permutations into a product of transpositions:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}, \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 4 & 1 & 2 & 5 & 8 & 3 & 7 \end{pmatrix}, \quad \rho = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 4 & 6 & 1 & 10 & 2 & 5 & 9 & 8 & 3 & 7 \end{pmatrix}.$$

○

**2.E.18.** Determine the parity of the following permutations:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 5 & 6 & 4 & 1 & 2 & 3 \end{pmatrix}, \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 6 & 7 & 1 & 2 & 3 & 8 & 4 & 5 \end{pmatrix}, \quad \rho = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 9 & 7 & 1 & 10 & 2 & 5 & 4 & 9 & 3 & 6 \end{pmatrix}.$$

○

**2.E.19.** For the matrix  $A$  shown below, determine all values of the complex parameter  $a \in \mathbb{C}$  that satisfy  $\det(A) = 1$ :

$$A = \begin{pmatrix} a & 1 & 1 & 1 \\ 0 & a & 1 & 1 \\ 0 & 1 & a & 1 \\ 0 & 0 & 0 & -a \end{pmatrix}.$$

**Solution.** Let us compute the determinant of  $A$  by expanding the first column of the matrix:

$$\det(A) = \begin{vmatrix} a & 1 & 1 & 1 \\ 0 & a & 1 & 1 \\ 0 & 1 & a & 1 \\ 0 & 0 & 0 & -a \end{vmatrix} = a \cdot \begin{vmatrix} a & 1 & 1 \\ 1 & a & 1 \\ 0 & 0 & -a \end{vmatrix} = -a^4 + a^2.$$

In Sage we can verify this computation by the cell

```
var("a")
A=matrix([[a, 1, 1, 1], [0, a, 1, 1], [0, 1, a, 1], [0, 0, 0, -a]]); det(A)
```

Thus, the equation  $\det(A) = 1$  is equivalent to  $a^4 - a^2 + 1 = 0$ . Substituting  $t = a^2$  we obtain  $t^2 - t + 1$  with roots

$$t_1 = \frac{1 + i\sqrt{3}}{2} = \cos(\pi/3) + i \sin(\pi/3), \quad t_2 = \frac{1 - i\sqrt{3}}{2} = \cos(\pi/3) - i \sin(\pi/3) = \cos(-\pi/3) + i \sin(-\pi/3).$$

Therefore, there are four possible values for the parameter  $a$ :

$$a_1 = \cos(\pi/6) + i \sin(\pi/6) = \sqrt{3}/2 + i/2, \quad a_2 = \cos(7\pi/6) + i \sin(7\pi/6) = -\sqrt{3}/2 - i/2, \\ a_3 = \cos(-\pi/6) + i \sin(-\pi/6) = \sqrt{3}/2 - i/2, \quad a_4 = \cos(5\pi/6) + i \sin(5\pi/6) = -\sqrt{3}/2 + i/2.$$

Of course, solving the equation  $\det(A) = 1$  with Sage is straightforward. Simply add the code `solve(det(A) == 1, a)` to the previous cell. This provides a fast way to verify the result.

Note that there is an alternative way to specify the possible values of  $a$ . For this, multiply the equation  $\det(A) = 1$ , i.e., the equation  $a^4 - a^2 + 1 = 0$  by  $a^2 + 1$ . This gives  $0 = (a^2 + 1)(a^4 - a^2 + 1) = a^6 + 1$ , and the idea is that we can easier treat the equation  $a^6 + 1$ , comparing the original equation. Indeed, recall by 1.G.17 that the equation  $a^6 = -1$  has six complex solutions, which can be presented as

$$a = \cos \varphi + i \sin \varphi, \quad \text{where } \varphi = \pi/6 + k\pi/3 = (2k + 1)\pi/6, \quad k = 0, 1, 2, 3, 4, 5.$$

Between them, we must discard the two choices  $k = 1$  and  $k = 4$ , since these choices solve  $a^2 + 1 = 0$ , and not our equation  $a^4 - a^2 + 1 = 0$ . Hence we conclude that  $a = \cos \varphi + i \sin \varphi$  with  $\varphi = (2k + 1)\pi/6$ , for  $k = 0, 2, 3$ , and  $k = 5$ .  $\square$

**2.E.20.** Establish whether the following matrix is invertible:

$$A = \begin{pmatrix} 3 & 2 & -1 & 2 \\ 4 & 1 & 2 & -4 \\ -2 & 2 & 4 & 1 \\ 2 & 3 & -4 & 8 \end{pmatrix}.$$

**Solution.** We just need to compute  $\det(A)$ . We will apply Laplace theorem (see 2.2.9) by expanding the first row of the given matrix. This gives

$$\begin{aligned} \det(A) &= 3 \cdot \begin{vmatrix} 1 & 2 & -4 \\ 2 & 4 & 1 \\ 3 & -4 & 8 \end{vmatrix} - 2 \cdot \begin{vmatrix} 4 & 2 & -4 \\ -2 & 4 & 1 \\ 2 & -4 & 8 \end{vmatrix} + (-1) \cdot \begin{vmatrix} 4 & 1 & -4 \\ -2 & 2 & 1 \\ 2 & 3 & 8 \end{vmatrix} - 2 \cdot \begin{vmatrix} 4 & 1 & 2 \\ -2 & 2 & 4 \\ 2 & 3 & -4 \end{vmatrix} \\ &= 3 \cdot 90 - 2 \cdot 180 - 1 \cdot 110 - 2 \cdot (-100) = 0. \end{aligned}$$

Thus  $\det(A) = 0$  and  $A$  cannot be invertible. As a verification by Sage, give the cell

```
A=matrix([[3, 2, -1, 2], [4, 1, 2, -4], [-2, 2, 4, 1], [2, 3, -4, 8]])
det(A)
```

Note that giving the command `A.inverse()`, Sage will return an error, since  $A$  is singular, i.e.,  $\det(A) = 0$ .  $\square$



We now present the determinant of the famous “Vandermonde matrix”. The Vandermonde matrix  $V_n = (V_{ij})$  is the square matrix of size  $n$ , with columns formed by the powers of a given vector  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ , i.e.,  $V_{ij} = x_i^{j-1}$  for  $i, j = 1, \dots, n$ , see below. This matrix is fundamental in both pure and applied mathematics. For example, in Chapter 5, we will explore its role in polynomial interpolation (see 5.1.5). Beyond mathematics, Vandermonde matrices are also significant in the natural sciences, particularly in economics and statistics.

**2.E.21. Vandermonde determinant.** The matrix

$$V_n := \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2^{n-1} & \dots & x_n^{n-1} \end{pmatrix},$$

where  $x_1, \dots, x_n \in \mathbb{R}$ , is the so called Vandermonde matrix. Prove that  $\det(V_n) = \prod_{1 \leq i < j \leq n} (x_j - x_i)$ .

**Solution.** The determinant will be a polynomial on  $x_1, \dots, x_n$ , which we may denote by  $V_n(x_1, \dots, x_n) = \det(V_n)$ . Notice for  $n = 2$  the claim holds trivially. We will proceed by induction on  $n$ . In fact we can work with the transposed Vandermonde matrix  $V_n^T$ , which satisfies

$$\det(V_n^T) = \det(V_n) = V_n(x_1, \dots, x_n), \quad n = 1, 2, \dots$$

By subtracting the first row from all other rows and then expanding the first column we obtain

$$V_n(x_1, x_2, \dots, x_n) = \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 0 & x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-1} - x_1^{n-1} \end{vmatrix} = \begin{vmatrix} x_2 - x_1 & x_2^2 - x_1^2 & \dots & x_2^{n-1} - x_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_n - x_1 & x_n^2 - x_1^2 & \dots & x_n^{n-1} - x_1^{n-1} \end{vmatrix}.$$

Let us now take out  $x_{i+1} - x_1$  from the  $i$ -th row, for all  $i \in \{1, 2, \dots, n-1\}$ . Then we deduce that

$$V_n(x_1, x_2, \dots, x_n) = (x_2 - x_1) \cdots (x_n - x_1) \cdot \begin{vmatrix} 1 & x_2 + x_1 & \dots & \sum_{j=0}^{n-2} x_2^{n-j-2} x_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & \dots & \sum_{j=0}^{n-2} x_n^{n-j-2} x_1^j \end{vmatrix}.$$

By subtracting from every column the  $x_1$ -multiple of the previous column, starting with the last and ending with the second one, we obtain

$$\begin{vmatrix} 1 & x_2 + x_1 & \cdots & \sum_{j=0}^{n-2} x_2^{n-j-2} x_1^j \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n + x_1 & \cdots & \sum_{j=0}^{n-2} x_n^{n-j-2} x_1^j \end{vmatrix} = \begin{vmatrix} 1 & x_2 & \cdots & x_2^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^{n-2} \end{vmatrix}.$$

Therefore, if for simplicity set  $V_n = V_n(x_1, x_2, \dots, x_n)$ , we see that  $V_n = (x_2 - x_1) \cdots (x_n - x_1) \cdot V_{n-1}(x_2, \dots, x_n)$ . Now, it is clear that  $V_2(x_{n-1}, x_n) = x_n - x_{n-1}$ , and thus by induction it follows that  $V_n(x_1, x_2, \dots, x_n) = \prod_{1 \leq i < j \leq n} (x_j - x_i)$ .

Observe that the determinant is non-zero, whenever the numbers  $x_1, \dots, x_n$  are mutually distinct. For an alternative proof of this classical result we refer to 5.1.5.  $\square$

**2.E.22. Vandermonde matrix in Sage.** Compute in Sage the determinant of a  $3 \times 3$  Vandermonde matrix  $V_3(x_1, x_2, x_3)$  using appropriately the `matrix.vandermonde` command.

**Solution.** To use the `matrix.vandermonde` command in Sage we need to specify first a vector. Since the task requires a general  $3 \times 3$  Vandermonde matrix, this vector should be of size  $3 \times 1$  and symbolic. Note also that the `matrix.vandermonde` command returns the transpose of the Vandermonde matrix. Hence, for  $n = 3$  one can use the following code block:

```
var("x1", "x2", "x3")
x=vector([x1, x2, x3])
V3=matrix.vandermonde(x, ring=None)
show(V3.T)
show("The determinant of the Vandermonde matrix is given by:", det(V3.T).factor())
```

Sage generates the desired matrix  $V_3(x_1, x_2, x_3)$ ,

$$\begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{pmatrix},$$

and also provides the result for its determinant, as shown below:

The determinant of the Vandermonde matrix is given by:  $-(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$ .

An alternative goes as follows:

```
A=matrix.vandermonde(SR.var(["x1", "x2", "x3"]))
show(A.T); show(det(A.T).factor())
```

$\square$

**2.E.23.** Use two different methods to solve the linear system

$$\begin{cases} x_1 + x_2 + x_3 + x_4 = 2, \\ x_1 + x_2 - x_3 - x_4 = 3, \\ x_1 - x_2 + x_3 - x_4 = 3, \\ x_1 - x_2 - x_3 + x_4 = 5. \end{cases}$$

Hints: Solve the system by computing the inverse of the coefficient matrix. Another method is based on the Cramer's rule.

**Solution.** In matrix notation we have  $Ax = b$ , where

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 5 \end{pmatrix},$$

respectively. We compute  $\det(A) = -16$ , hence  $A$  is invertible and the linear system has a unique solution which occurs by multiplying the matrix equation  $Ax = b$  by  $A^{-1}$  from the left, i.e.,  $x = A^{-1}b$ . Thus, it suffices to compute the inverse of  $A$  and next do the matrix multiplication  $A^{-1}b$ . We see that

$$A^{-1} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{pmatrix} \implies x = A^{-1}b = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} \frac{13}{4} \\ -\frac{3}{4} \\ -\frac{3}{4} \\ \frac{1}{4} \end{pmatrix}.$$

Thus, the solution has the following form:

$$\left(\frac{13}{4}, -\frac{3}{4}, -\frac{3}{4}, \frac{1}{4}\right)^T \in \mathbb{R}^4.$$

In order to compute  $A^{-1}$  we applied the relation  $A^{-1} = (1/\det(A))\text{adj}(A)$ . Recall that the algebraic adjoint matrix  $\text{adj}(A)$  of  $A$  is defined by  $\text{adj}(A) = (A_{ij})^T$ , where  $A_{ij}$  is the algebraic complement of the element  $a_{ij}$  of the matrix  $A$ , that is,  $A_{ij} = (-1)^{i+j} \det(\hat{A}_{ij})$ . Here,  $\hat{A}_{ij}$  denotes the matrix obtained by  $A$  by deleting the  $i$ th row and  $j$ th column. We compute

$$\begin{array}{l}
 A_{11} = (-1)^{1+1} \det \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{12} = (-1)^{1+2} \det \begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{13} = (-1)^{1+3} \det \begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{14} = (-1)^{1+4} \det \begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} = -4, \\
 A_{21} = (-1)^{2+1} \det \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{22} = (-1)^{2+2} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{23} = (-1)^{2+3} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = 4, \\
 A_{24} = (-1)^{2+4} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} = 4, \\
 A_{31} = (-1)^{3+1} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{32} = (-1)^{3+2} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = 4, \\
 A_{33} = (-1)^{3+3} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = -4, \\
 A_{34} = (-1)^{3+4} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} = 4, \\
 A_{41} = (-1)^{4+1} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \end{pmatrix} = -4, \\
 A_{42} = (-1)^{4+2} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \end{pmatrix} = 4, \\
 A_{43} = (-1)^{4+3} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix} = 4, \\
 A_{44} = (-1)^{4+4} \det \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} = -4.
 \end{array}$$

Hence by applying the rule  $A^{-1} = (1/\det(A))\text{adj}(A)$  we obtain the given expression for the inverse  $A^{-1}$  of  $A$ .  $\square$

**2.E.24.** Confirm all the computations presented in 2.E.23 via Sage.  $\circ$

**2.E.25.** Present in Sage the solution given in 2.E.23 via the Cramer's rule.  $\circ$

**2.E.26.** Explain why the matrix

$$A = \begin{pmatrix} 5 & 0 & 8 & 0 \\ 1 & 5 & 0 & 4 \\ 4 & 0 & 5 & 0 \\ 0 & 7 & 0 & 8 \end{pmatrix}$$

is invertible. Next use Sage to compute its algebraic adjoint and its inverse.

**Solution.** Computing the determinant of  $A$  we see that  $\det(A) = -84$ . Thus  $A$  has an inverse, with an inverse given by  $A^{-1} = \frac{1}{\det(A)} \cdot \text{adj}(A)$ . As we know, in order to derive the adjoint and the inverse of  $A$  in Sage, the corresponding code has the form

```
A=matrix([[5, 0, 8, 0], [1, 5, 0, 4], [4, 0, 5, 0], [0, 7, 0, 8]])
print(det(A)); show(A.adjugate()); show(A.inverse())
```

Executing this block we get the following expressions:

$$\text{adj}(A) = \begin{pmatrix} 60 & 0 & -96 & 0 \\ -40 & -56 & 64 & 28 \\ -48 & 0 & 60 & 0 \\ 35 & 49 & -56 & -35 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} -\frac{5}{7} & 0 & \frac{8}{7} & 0 \\ \frac{10}{21} & \frac{2}{3} & -\frac{16}{21} & -\frac{1}{3} \\ \frac{4}{7} & 0 & -\frac{5}{7} & 0 \\ -\frac{5}{12} & -\frac{7}{12} & \frac{2}{3} & \frac{5}{12} \end{pmatrix}.$$

**2.E.27.** For the matrix  $F = \begin{pmatrix} \alpha & \beta & 0 \\ \gamma & \delta & 0 \\ 0 & 0 & 1 \end{pmatrix}$ , where  $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ , compute its algebraically adjoint matrix  $\text{adj}(F)$ . □

**2.E.28.** Calculate the algebraically adjoint matrix for the matrices

$$A = \begin{pmatrix} 3 & -2 & 0 & -1 \\ 0 & 2 & 2 & 1 \\ 1 & -2 & -3 & -2 \\ 0 & 1 & 2 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1+i & 2i \\ 3-2i & 6 \end{pmatrix},$$

where as usual  $i$  denotes the imaginary unit. □

**2.E.29.** Show that the matrix

$$A = \begin{pmatrix} 2 & -3 & -5 \\ -1 & 4 & 5 \\ 1 & -3 & -4 \end{pmatrix} \in \text{Mat}_3(\mathbb{R})$$

is *idempotent*, i.e.,  $A^2 = A$ .<sup>13</sup> Then, based explicitly on this property of  $A$ , compute  $AB$  and  $B^2$ , where  $B = A - E$ , with  $E$  being the  $3 \times 3$  identity matrix. Moreover, compute the determinant  $\det(A)$  of  $A$ . □

**2.E.30.** Let  $A, B, P \in \text{Mat}_n(\mathbb{R})$  be invertible  $n \times n$  matrices such that  $B = PAP^{-1}$ . Can we conclude that  $\det(A) = \det(B)$  and  $\det(A^{-1}B) = 1$ ? Use the Cauchy theorem (see 2.2.7) to support your answer. □

### C) MATERIAL ON VECTOR SPACES AND LINEAR MAPPINGS

**2.E.31.** Consider the complex numbers  $\mathbb{C}$  as a real vector space. Determine the coordinates of  $z = 2 + i$  in the basis given by the roots of the polynomial  $x^2 + x + 1$ .

**Solution.** Because roots of the given polynomial are  $-\frac{1}{2} + i\frac{\sqrt{3}}{2}$  and  $-\frac{1}{2} - i\frac{\sqrt{3}}{2}$ , we have to determine the coordinates  $(a, b)$  of the vector  $z = 2 + i$  in the basis  $(-\frac{1}{2} + i\frac{\sqrt{3}}{2}, -\frac{1}{2} - i\frac{\sqrt{3}}{2})$ . These real numbers  $a, b$  are uniquely determined by the condition

$$a \cdot \left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right) + b \cdot \left(-\frac{1}{2} - i\frac{\sqrt{3}}{2}\right) = 2 + i.$$

By equating separately the real and the imaginary parts of the equation, we obtain a system of two linear equations in two variables:

$$\left\{ \begin{array}{l} -\frac{1}{2}a - \frac{1}{2}b = 2, \\ \frac{\sqrt{3}}{2}a - \frac{\sqrt{3}}{2}b = 1 \end{array} \right\}.$$

The solution has the form  $a = -2 + \frac{\sqrt{3}}{3}$ , and  $b = -2 - \frac{\sqrt{3}}{3}$ , respectively, hence the required coordinates are given by  $(-2 + \frac{1}{\sqrt{3}}, -2 - \frac{1}{\sqrt{3}})$ . □

**2.E.32.** Consider the complex numbers  $\mathbb{C}$  as a real vector space. Determine the coordinates of the number  $2 + i$  in the basis given by the roots of the polynomial  $x^2 - x + 1$ . □

**2.E.33.** Express the vector  $(5, 1, 11)^T \in \mathbb{R}^3$ , as a linear combination of the vectors  $(3, 2, 2)^T$ ,  $(2, 3, 1)^T$ , and  $(1, 1, 3)^T$ . □

**2.E.34.** Given the vectors  $u_1 = (1, 1, a, 1)^T$ ,  $u_2 = (1, b, 1, 1)^T$ ,  $u_3 = (c, 1, 1, 1)^T$  in  $\mathbb{R}^4$ , specify the values of the parameters  $a, b, c \in \mathbb{R}$ , such that  $u_1, u_2, u_3$  are linearly dependent. □

**2.E.35.** Let  $V$  be vector space with a basis formed by the vectors  $u, v, w, z$ . Determine whether or not the following vectors of  $V$  are linearly independent:  $u - 3v + z$ ,  $v - 5w - z$ ,  $3w - 7z$  and  $u - w + z$ . □

**2.E.36.** Complete the vectors  $1 - x^2 + x^3$ ,  $1 + x^2 + x^3$ ,  $1 - x - x^3$  to a basis of the vector space  $\mathbb{R}_3[x]$  of polynomials over  $\mathbb{R}$  of degree at most 3. □

**2.E.37.** Depending on the parameter  $t \in \mathbb{R}$ , determine the dimension of the subspace  $U \subset \mathbb{R}^3$  for the following two cases:

(a) The subspace  $U$  is generated by the vectors  $u_1 = (1, 1, 1)^T$ ,  $u_2 = (1, t, 1)^T$ , and  $u_3 = (2, 2, t)^T$ ;

(b) The subspace  $U$  is generated by the vectors  $u_1 = (t, t, t)^T$ ,  $u_2 = (-4t, -4t, 4t)^T$ , and  $u_3 = (-2, -2, -2)^T$ . □

<sup>13</sup>More general, a  $n \times n$  matrix  $A$  is called *idempotent* if  $A^2 = A$ . Idempotent matrices are also called *projectors* (or *projections*), see also the exercise 2.E.61 below.

**2.E.38.** (a) Using a similar approach to the proof shown in 2.C.31, prove that the linear mapping

$$\begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \mapsto (a, b, c, d, e, f)^T,$$

is a linear isomorphism between the vector spaces  $\text{Mat}_{3,2}(\mathbb{R})$  and  $\mathbb{R}^6$ . Next find a basis of  $\text{Mat}_{3,2}(\mathbb{R})$ .

(b) Determine a basis of the subspace

$$U = \text{span}_{\mathbb{R}} \left\{ U_1 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}, U_2 = \begin{pmatrix} 0 & 1 \\ 2 & 3 \\ 4 & 5 \end{pmatrix}, U_3 = \begin{pmatrix} -1 & 0 \\ 1 & 2 \\ 3 & 4 \end{pmatrix}, U_4 = \begin{pmatrix} -2 & -1 \\ 0 & 1 \\ 2 & 3 \end{pmatrix} \right\} \subset \text{Mat}_{3,2}(\mathbb{R}).$$

Extend this basis to a basis of the whole space  $\text{Mat}_{3,2}(\mathbb{R})$  of real  $3 \times 2$  matrices.

**Solution.** (a) An explicit proof here is very similar to 2.C.31, so you can easily complete the necessary steps alone. As for a basis, verify (also in Sage, with the same method described in 2.C.31) that the following matrices

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad E_5 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad E_6 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix},$$

are linearly independent, and hence they form a basis of the vector space  $\text{Mat}_{3,2}(\mathbb{R})$  of all real  $3 \times 2$  matrices.

(b) Let us use the indicated isomorphism and consider real  $3 \times 2$  matrices as vectors in  $\mathbb{R}^6$ . In this way, the four given matrices can be identified with the rows of the following matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ -1 & 0 & 1 & 2 & 3 & 4 \\ -2 & -1 & 0 & 1 & 2 & 3 \end{pmatrix}.$$

It is easy to show that  $A$  has rank 2. This implies that the subspace  $U$  is generated just by the first two matrices  $U_1, U_2$ , which consequently form a basis for  $U$ . Hence in fact the rest two matrices  $U_3, U_4 \in U$  should be linear combinations of  $U_1, U_2$ , and indeed you should be able to verify that  $U_3 = -U_1 + 2U_2$ , and  $U_4 = -2U_1 + 3U_2$ , respectively.

Now, there are many options for extending this basis to a basis of the whole space. For instance, we may choose the first two of the given matrices above, together with the matrices  $E_3, E_4, E_5, E_6$  of the basis mentioned in (a). In order to prove the linear independence we may first apply the isomorphism established in (a), and then compute the determinant of the matrix induced by  $U_1, U_2, E_3, E_4, E_5, E_6$ . We see that



$$\begin{vmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} = 1 \neq 0.$$

Clearly, the dimension is 6, so the linear spanning is automatic and we have obtained a basis. Prove that the choice of the matrices  $E_3, E_4, E_5, E_6$  is not so crucial, in the sense that we could fix any four matrices between  $E_1, \dots, E_6$ .  $\square$

**2.E.39.** Consider the linear subspaces  $U$  and  $V$  of  $\mathbb{R}^3$ , generated by the vectors  $u_1 = (1, 1, -3)^T$ ,  $u_2 = (1, 2, 2)^T$ , and  $v_1 = (1, 1, -1)^T$ ,  $v_2 = (1, 2, 1)^T$ ,  $v_3 = (1, 3, 3)^T$ , respectively. Determine the intersection  $U \cap V$ .  $\circ$

**2.E.40.** In the vector space  $\mathbb{R}^3$ , determine the matrix of the orthogonal projection onto the plane  $x + y - 2z = 0$ .  $\circ$

**2.E.41.** In the vector space  $\mathbb{R}^3$ , determine the matrix of the orthogonal projection on the plane  $2x - y + 2z = 0$ .  $\circ$

**2.E.42.** Consider the following subspaces of  $\mathbb{R}^4$ :

$$U = \text{span}_{\mathbb{R}} \{(2, 1, 2, 2)^T\}, \quad W = \text{span}_{\mathbb{R}} \{(-1, 0, -1, 2)^T, (-1, 0, 1, 0)^T, (0, 0, 1, -1)^T\}.$$

(a) Is it true that  $U \perp W$  with respect to the dot product on  $\mathbb{R}^4$ ?

(b) Is it true that  $U^\perp = W$ , and hence  $\mathbb{R}^4 = U \oplus W$ ?  $\circ$

**2.E.43.** Let  $W$  be the subspace of the Euclidean space  $(\mathbb{R}^4, \langle \cdot, \cdot \rangle)$  spanned by the vectors  $E_1 = (1, 0, 0, 1)^T$ ,  $E_2 = (0, 1, 0, 1)^T$  and  $E_3 = (1, 2, 1, 0)^T$ . Present an orthonormal basis of  $W$ . Next deduce that its orthogonal complement  $W^\perp$  with respect to the standard Euclidean product  $\langle \cdot, \cdot \rangle$  is one-dimensional and find a generator.  $\circ$



**2.E.44.** Consider the linear mapping  $F : \mathbb{R}^4 \rightarrow \mathbb{R}^2$  given by



$$\mathbb{R}^4 \ni u = (x_1, x_2, x_3, x_4)^T \mapsto F(u) = \begin{pmatrix} x_1 + 2x_3 + 3x_4 \\ x_2 + 4x_3 + 5x_4 \end{pmatrix} \in \mathbb{R}^2.$$

- Find the matrix  $A$  of  $F$  with respect to the standard bases on  $\mathbb{R}^4$  and  $\mathbb{R}^2$ .
- Find the column space, row space, kernel and cokernel of the matrix of  $F$ , and next use Sage to confirm your answers by applying the commands presented in 2.C.20.
- Deduce that the transpose operator of  $F$  is injective.
- Prove the orthogonal decompositions

$$\mathbb{R}^4 = C(A^T) \oplus \text{Ker}(A), \quad \mathbb{R}^2 = C(A) \oplus \text{Ker}(A^T) = C(A) \oplus \{0\} \cong C(A).$$

**Solution.** (i) Let us denote the vectors of the standard bases of  $\mathbb{R}^4$  and  $\mathbb{R}^2$  by  $e_1 = (1, 0, 0, 0)^T$ ,  $e_2 = (0, 1, 0, 0)^T$ ,  $e_3 = (0, 0, 1, 0)^T$ ,  $e_4 = (0, 0, 0, 1)^T$  and  $\alpha_1 = (1, 0)^T$ ,  $\alpha_2 = (0, 1)^T$ , respectively. Then, we see that

$$\begin{aligned} F(e_1) &= (1, 0)^T = \alpha_1, \\ F(e_2) &= (0, 1)^T = \alpha_2, \\ F(e_3) &= (2, 4)^T = 2\alpha_1 + 4\alpha_2, \\ F(e_4) &= (3, 5)^T = 3\alpha_1 + 5\alpha_2. \end{aligned}$$

Thus, the matrix of  $F$  with respect to the standard bases is given by  $A = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 4 & 5 \end{pmatrix}$ . We may verify this result in Sage, by applying the method described in 2.C.28, for example:

```
V=RR^4
W=RR^2;
var("x1, x2, x3, x4")
f(x1, x2, x3, x4)=[x1+2*x3+3*x4, x2+4*x3+5*x4]
T=linear_transformation(V, W, f)
A=T.matrix(side="right");
show(A)
```

(ii) The matrix  $A$  is already in RREF, and hence its rank equals two,  $\text{rank}(A) = 2$ . Here is a confirmation via Sage:

```
A=matrix([[1, 0, 2, 3], [0, 1, 4, 5]])
print(A==A.rref()); print(rank(A))
```

In particular, the rows of  $A$  are linearly independent. Hence, the row space of  $A$  is a 2-dimensional subspace of  $\mathbb{R}^4$  given by  $C(A^T) = \text{span}_{\mathbb{R}} \{(1, 0, 2, 3)^T, (0, 1, 4, 5)^T\}$ . Let us denote by

$$\alpha_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \alpha_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \alpha_3 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \quad \alpha_4 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

the column vectors of  $A$ . Then, we see that  $\alpha_3 = 2\alpha_1 + 4\alpha_2$ ,  $\alpha_4 = 3\alpha_1 + 5\alpha_2$ , while  $\alpha_1, \alpha_2$  are linearly independent. Therefore,  $C(A)$  is a 2-dimensional subspace of  $\mathbb{R}^2$  generated by the vectors of the standard basis of  $\mathbb{R}^2$ , i.e.,  $C(A) = \text{span}_{\mathbb{R}} \{(1, 0)^T, (0, 1)^T\}$ .

The kernel of  $A$  corresponds to solutions of the homogeneous system  $Au = 0$ , i.e.,

$$\begin{pmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 4 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This corresponds to the system  $\{x_1 + 2x_3 + 3x_4 = 0, x_2 + 4x_3 + 5x_4 = 0\}$  and we deduce that  $\text{Ker}(A)$  is a 2-dimensional subspace of  $\mathbb{R}^4$  given by

$$\text{Ker}(A) = \left\{ \begin{pmatrix} -2t - 3s \\ -4t - 5s \\ t \\ s \end{pmatrix} : t, s \in \mathbb{R} \right\} = \text{span}_{\mathbb{R}} \left\{ \begin{pmatrix} -2 \\ -4 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -3 \\ -5 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

The left null space of  $A$  corresponds to solutions of the homogeneous system  $A^T w = 0$  for some  $w = (y_1, y_2)^T \in \mathbb{R}^2$ , i.e.,

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 4 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

This system has a unique solution given by  $y_1 = y_2 = 0$ . Thus, the cokernel  $\text{Ker}(A^T)$  is trivial.

To verify these results in Sage, one may proceed as in 2.C.20:

```
A=matrix([[1, 0, 2, 3], [0, 1, 4, 5]])
print("\nThe column space of A is given by:", A.column_space())
print("\nThe row space of A is given by:", A.row_space())
print("\nThe kernel of A is given by:", A.right_kernel())
print("\nThe cokernel of A is given by:", A.left_kernel())
```

For your convenience here is the output of Sage:

The column space of A is given by: Free module of degree 2 and rank 2 over Integer Ring  
Echelon basis matrix:

```
[1 0]
[0 1]
```

The row space of A is given by: Free module of degree 4 and rank 2 over Integer Ring  
Echelon basis matrix:

```
[1 0 2 3]
[0 1 4 5]
```

The kernel of A is given by: Free module of degree 4 and rank 2 over Integer Ring  
Echelon basis matrix:

```
[ 1  1  1 -1]
[ 0  2 -3  2]
```

The cokernel of A is given by: Free module of degree 2 and rank 0 over Integer Ring  
Echelon basis matrix:

```
[]
```

(iii) The transpose of  $F$  is the linear operator  $\hat{F}: \mathbb{R}^2 \rightarrow \mathbb{R}^4$  whose matrix presentation (with respect to the standard bases of  $\mathbb{R}^2$  and  $\mathbb{R}^4$ ) is given by the transpose of  $A$ . Thus  $\hat{F}(w) = A^T w$  for any  $w \in \mathbb{R}^2$  and  $\hat{F}$  is injective since  $\text{Ker}(A^T) = \{0\}$ .

(iv) Above we saw that  $C(A^T) = \text{span}_{\mathbb{R}}\{(1, 0, 2, 3)^T, (0, 1, 4, 5)^T\}$  and  $\text{Ker}(A) = \text{span}_{\mathbb{R}}\{(-2, -4, 1, 0)^T, (-3, -5, 0, 1)^T\}$ .

Suppose that  $v$  is a vector in  $C(A^T) \cap \text{Ker}(A)$ . Then we can think  $v$  as a linear combination of the vectors  $E_1 = (1, 0, 2, 3)^T$  and  $E_2 = (0, 1, 4, 5)^T$ , satisfying the additional condition  $Av = 0$ . This gives the homogeneous system

$$Av = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 4 & 5 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ 2\alpha + 4\beta \\ 3\alpha + 5\beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \iff \begin{pmatrix} 14\alpha + 23\beta \\ 23\alpha + 42\beta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which has the zero solution  $\alpha = \beta = 0$  as unique. Thus,  $v = 0$  and  $C(A^T) \cap \text{Ker}(A) = \{0\}$ . This, along with the relation  $\dim C(A^T) + \dim \text{Ker}(A) = 2 + 2 = 4 = \dim \mathbb{R}^4$  proved above, precisely defines the conditions for a direct sum, i.e.,  $\mathbb{R}^4 = C(A^T) \oplus \text{Ker}(A)$ . The orthogonality condition  $C(A^T) \perp \text{Ker}(A)$  follows by 2.C.56. For a direct verification, observe that for any  $u \in C(A^T)$  and  $w \in \text{Ker}(A)$  we have

$$u = aE_1 + vE_2, \quad w = cE_3 + dE_4$$

for some scalars  $a, b, c, d \in \mathbb{R}$ , where  $E_3 = (-2, -4, 1, 0)^T$ ,  $E_4 = (-3, -5, 0, 1)^T$  and the vectors  $E_1, E_2$  are as above. Thus, we only need prove that  $\langle E_1, E_3 \rangle = 0 = \langle E_1, E_4 \rangle$  and  $\langle E_2, E_3 \rangle = 0 = \langle E_2, E_4 \rangle$ , which is straightforward:

$$E_3^T E_1 = -2 + 2 = 0, \quad E_4^T E_1 = -3 + 3 = 0, \quad E_3^T E_2 = -4 + 4 = 0, \quad E_4^T E_2 = -5 + 5 = 0.$$

Recall that in Sage these scalar products can be computed using the `dot_product` command:

```
E1=vector([1, 0, 2, 3])
E2=vector([0, 1, 4, 5])
E3=vector([-2, -4, 1, 0])
E4=vector([-3, -5, 0, 1])
print(E1.dot_product(E3)==0)
print(E2.dot_product(E3)==0)
```

```
print (E1.dot_product (E4)==0)
print (E2.dot_product (E4)==0)
```

The second decomposition  $\mathbb{R}^2 = C(A) \oplus \text{Ker}(A^T)$  can be verified in an analogous way, but it is much easier and reduces to  $\mathbb{R}^2 \cong C(A)$ , as  $\text{Ker}(A^T)$  is trivial.  $\square$

**2.E.45. Remark.** Recall from the theoretical task 2.C.56 that any  $m \times n$  matrix  $A$  over  $\mathbb{R}$  satisfies

$$C(A) \perp \text{Ker}(A^T), \quad \text{and} \quad C(A^T) \perp \text{Ker}(A),$$

with respect to the corresponding dot products. In fact, it is not so difficult to prove that

$$C(A) \cap \text{Ker}(A^T) = \{0\}, \quad C(A^T) \cap \text{Ker}(A) = \{0\}, \quad \dim \text{Ker}(A^T) = m - \text{rank}(A), \quad \dim \text{Ker}(A) = n - \text{rank}(A),$$

for any such matrix  $A$ , and all these properties together lead to the following orthogonal direct sum decompositions:

$$\mathbb{R}^m = C(A) \oplus \text{Ker}(A^T), \quad \mathbb{R}^n = C(A^T) \oplus \text{Ker}(A).$$

Note that this result extends to complex matrices with appropriate adjustments (for instance one should use the standard Hermitian inner product on  $\mathbb{C}^m$  and  $\mathbb{C}^n$ , and replace  $A^T$  by the conjugate transpose of the complex matrix  $A$ , which we will mainly explore in Chapter 3, see also below).

**2.E.46.** Consider the linear mapping  $F : \mathbb{R}^4 \rightarrow \mathbb{R}^2$  presented in 2.E.44 and the sets

$$\begin{aligned} \underline{u} &= \{u_1 = (1, 0, 0, 0)^T, u_2 = (0, 1, 0, 0)^T, u_3 = (0, 0, 1, 0)^T, u_4 = (1, 1, 1, 1)^T\}, \\ \underline{v} &= \{v_1 = (1, 1)^T, v_2 = (2, -1)^T\}. \end{aligned}$$

- i) Use Sage to prove that  $\underline{u}$  and  $\underline{v}$  are bases of  $\mathbb{R}^4$  and  $\mathbb{R}^2$ , respectively.
- ii) Find the new coordinates  $w_{\underline{u}}$  of the vector  $w = (1, 2, 3, 4)^T \in \mathbb{R}^4$  with respect to  $\underline{u}$ .
- iii) Find the new matrix presentation of  $F$  with respect the new bases  $\underline{u}$  and  $\underline{v}$  presented above.
- iv) Verify that the image of  $w$  by  $F$  remains invariant across the two different matrix representations of  $F$  (this is true for any vector  $w \in \mathbb{R}^4$ ). In particular, demonstrate that  $F(w) = Q F(w)_{\underline{v}}$ , where  $F(w)_{\underline{v}}$  is the image of  $w$  in the new coordinates on  $\mathbb{R}^2$ , and  $Q$  is the transition matrix from the new basis  $\underline{v}$  to the standard basis of  $\mathbb{R}^2$ .

**Solution.** (i) We can use Sage via the method described in 2.C.10 to verify the linear independence of  $u_1, \dots, u_4$

```
u1=vector(QQ, [1, 0, 0, 0])
u2=vector(QQ, [0, 1, 0, 0])
u3=vector(QQ, [0, 0, 1, 0])
u4=vector(QQ, [1, 1, 1, 1])
U=RR^4; L=[v1, v2, v3, v4]
U.linear_dependence(L)==[]
```

Similarly one can check the linear independence of  $v_1, v_2$ . Since for both cases the vectors are linearly independent we have  $\text{span}\{u_1, \dots, u_4\} \cong \mathbb{R}^4$  and  $\text{span}\{v_1, v_2\} \cong \mathbb{R}^2$ , respectively, and the claim for the bases follows.

(ii) The transition matrix from the basis  $\underline{u}$  to the standard basis  $\underline{e}$  of  $\mathbb{R}^4$ , is given by the matrix  $T$  having as columns vectors the vectors  $u_1, \dots, u_4$ . The transition matrix from  $\underline{e}$  to  $\underline{u}$  is the inverse matrix  $T^{-1}$ , given here: (compute  $T^{-1}$  on your own)

$$T = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus, the new coordinates of the vector  $w$  in terms of  $\underline{u}$  are given by  $w_{\underline{u}} = T^{-1}w = (-3, -2, -1, 4)^T$ . In Sage this task can be solved as in 2.C.42, hence we leave the verification to the reader.

(iii) Here one should apply the method we learned in 2.C.44. In particular, by 2.E.44 we know that the matrix of the mapping  $F : \mathbb{R}^4 \rightarrow \mathbb{R}^2$  with respect to the standard bases is given by  $A = \begin{pmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 4 & 5 \end{pmatrix}$ . Therefore, the new matrix of  $F$  with respect to the bases  $\underline{u}$  and  $\underline{v}$  is given by  $A' = [F]_{\underline{u}, \underline{v}} = Q^{-1}AT$ , where  $Q^{-1}$  is the inverse of the transition matrix  $Q$  from the new basis  $\underline{v}$  to the standard basis of  $\mathbb{R}^2$ , and  $T$  is as above. Since  $Q = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}$ , we compute  $Q^{-1} = -\frac{1}{3} \begin{pmatrix} -1 & -2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & -1/3 \end{pmatrix}$ , and

$$A' = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & -1/3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 & 3 \\ 0 & 1 & 4 & 5 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 & 10/3 & 26/3 \\ 1/3 & -1/3 & -2/3 & -4/3 \end{pmatrix}.$$

Here is a verification of these computations by Sage:

```
A=matrix([[1, 0, 2, 3], [0, 1, 4, 5]])
T=matrix([[1, 0, 0, 1], [0, 1, 0, 1], [0, 0, 1, 1], [0, 0, 0, 1]])
Q=matrix([[1, 2], [1, -1]])
Qinv=Q.inverse(); show(Qinv*A*T)
(iv)
```

In Sage one can find the matrix of the linear mapping  $F$  with respect to the new bases  $\underline{u}$  and  $\underline{v}$ , by combining the functions  $T = \text{linear\_transformation}(U, V, F)$  and  $T.\text{restrict\_domain}(U1).\text{restrict\_codomain}(V1)$ . Within the program, we should define the spaces spanned by the bases  $\underline{u}$  and  $\underline{v}$ , which we may denote by  $U1$  and  $V1$ , respectively (of course, these spaces are isomorphic to  $\mathbb{R}^4$  and  $\mathbb{R}^2$ ). This can be done by the command  $U1 = U.\text{subspace\_with\_basis}(u)$ , for example, where  $U$  is the domain of  $F$ . Additionally, to simplify the matrix displays we can instruct Sage to work within the vectors space  $\mathbb{Q}^4$  and  $\mathbb{Q}^2$ , instead of  $\mathbb{R}^4$  and  $\mathbb{R}^2$ . This adjustment does not affect the final result. The implementation of these concepts is as follows:

```
U=QQ^4; u1=vector([1, 0, 0, 0]); u2=vector([0, 1, 0, 0])
u3=vector([0, 0, 1, 0]); u4=vector([1, 1, 1, 1]); u=[u1, u2, u3, u4]
U1=U.subspace_with_basis(u)
V=QQ^2; v1=vector([1, 1]); v2=vector([2, -1]); v=[v1, v2]
V1=V.subspace_with_basis(v)
var("x1, x2, x3, x4")
F(x1, x2, x3, x4)=[x1+2*x3+3*x4, x2+4*x3+5*x4]
T=linear_transformation(U, V, F)
S=T.restrict_domain(U1).restrict_codomain(V1)
Ahat=S.matrix(side="right")
show(Ahat)
```

Check on your own that executing this code block Sage returns the matrix  $\begin{pmatrix} 1/3 & 2/3 & 10/3 & 26/3 \\ 1/3 & -1/3 & -2/3 & -4/3 \end{pmatrix}$ .

(iv) Above we have seen that the new coordinates of  $w \in \mathbb{R}^4$  in the basis  $\underline{u}$  are given by  $w_{\underline{u}} = T^{-1}w$ . On the other hand the matrix  $A' = [F]_{\underline{u}, \underline{v}}$  in the new bases is given by  $A' = Q^{-1}AT$ . The image  $F(w)$  in the new coordinates and in terms of matrices is the following vector:

$$F(w)_{\underline{v}} = A'w_{\underline{u}} = (Q^{-1}AT)(T^{-1}w) = Q^{-1}ATT^{-1}w = Q^{-1}Aw \in \mathbb{R}^2.$$

To bring this vector back to the original coordinates in  $\mathbb{R}^2$  we need to apply the transition matrix  $Q$ , which gives the result:  $QF(w)_{\underline{v}} = QQ^{-1}Aw = Aw = F(w)$ . Use the explicit expressions of  $w_{\underline{u}}$ ,  $A'$  and  $Q$  to confirm the claim, as well.  $\square$

**2.E.47.** Construct an orthogonal basis of the subspace  $W = \text{span}_{\mathbb{R}} \{(1, 1, 1, 1)^T, (1, 1, 1, -1)^T, (-1, 1, 1, 1)^T\} \subset \mathbb{R}^4$ .  $\circ$

**2.E.48.** In  $\mathbb{R}^4$ , find an orthogonal basis of the subspace of all linear combinations of the vectors  $(1, 0, 1, 0)^T$ ,  $(0, 1, 0, -7)^T$ , and  $(4, -2, 4, 14)^T$ . Moreover, find an orthogonal basis of the subspace generated by the vectors  $(1, 2, 2, -1)^T$ ,  $(1, 1, -5, 3)^T$  and  $(3, 2, 8, -7)^T$ .  $\circ$

**2.E.49.** Consider the following vectors of  $\mathbb{R}^5$ :  $u_1 = (1, 1, 2, 0, 0)^T$ ,  $u_2 = (1, -1, 0, 1, a)^T$  and  $u_3 = (1, b, 2, 3, -2)^T$ . Specify the values of the parameters  $a, b \in \mathbb{R}$  such that these vectors are pairwise orthogonal.  $\circ$

**2.E.50.** Let  $V$  the linear subspace of  $\mathbb{R}^4$  generated by the vectors  $(-1, 2, 0, 1)^T$ ,  $(3, 1, -2, 4)^T$ ,  $(-4, 1, 2, -4)^T$ , and  $(2, 3, -2, 5)^T$ . Determine the orthogonal complement of  $V$  with respect to the standard dot product in  $\mathbb{R}^4$ .  $\circ$

**2.E.51.** Consider the following two vectors  $u_1 = (1, -2, 2, 1)^T$  and  $u_2 = (1, 3, 2, 1)^T$  of  $\mathbb{R}^4$ . Extend their set into an orthogonal basis of the whole  $\mathbb{R}^4$ , with respect to the usual dot product.  $\circ$

**2.E.52.** Let  $V \subset \mathbb{R}^5$  the linear subspace generated by the vectors  $(1, 1, -1, -1, 0)^T$ ,  $(1, -1, -1, 0, -1)^T$ ,  $(1, 1, 0, 1, 1)^T$ , and  $(-1, 0, -1, 1, 1)^T$ . Find a basis for its orthogonal complement  $V^\perp$  with respect to the usual dot product.  $\circ$

**2.E.53.** In the Euclidean space  $\mathbb{R}^5$ , determine the orthogonal complement  $W^\perp$  of the subspace  $W$ , if

(a)  $W = \{(r + s + t, -r + t, r + s, -t, s + t); r, s, t \in \mathbb{R}\}$ ;

(b)  $W$  is the set of the solutions of the following system of linear equations  $\{x_1 - x_3 = 0, x_1 - x_2 + x_3 - x_4 + x_5 = 0\}$ .  $\circ$

So far we have primarily focused on real matrices, though manipulating complex matrices is quite similar. When dealing with complex matrices, a key operation is the complex conjugation. For a complex matrix  $A = (a_{ij}) \in \text{Mat}_{m,n}(\mathbb{C})$ , the complex conjugate  $\bar{A}$  is defined by  $\bar{A} = (\bar{a}_{ij})$ , where  $\bar{z}$  denotes the complex conjugate of  $z \in \mathbb{C}$ . Note that the matrix  $\bar{A}$  belongs to  $\text{Mat}_{m,n}(\mathbb{C})$ , as well. Another important operation is the conjugate transpose of a complex matrix, denoted by  $A^* = \overline{A^T} = \bar{A}^T$ . This results in a matrix in  $\text{Mat}_{n,m}(\mathbb{C})$ , provided that  $A \in \text{Mat}_{m,n}(\mathbb{C})$ . Any real matrix  $A$  can be viewed as a complex matrix with zero imaginary part, and hence such a matrix satisfies  $A^* = A^T$ . In Chapter 3 we will examine these concepts in greater detail. Hence you may consider the rest of this section as preparatory groundwork.



**2.E.54.** Recall that real matrix  $A \in \text{Mat}_n(\mathbb{R})$  is called symmetric if  $A = A^T$ , and skew-symmetric (or antisymmetric) if  $A = -A^T$ . Using the identity  $(A^T)^T = A$ , show that any square matrix  $A$  can be expressed as the sum of a symmetric and a skew-symmetric matrix. Then illustrate the statement using the matrix  $A$  given below and verify your answer in Sage:

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 6 & 3 & 0 \\ 2 & 2 & 4 \end{pmatrix}.$$

**2.E.55.** (a) Given a squared matrix  $A \in \text{Mat}_n(\mathbb{R})$ , show that  $A + A^T$  is symmetric, and that  $A - A^T$  is skew-symmetric.

(b) Given a  $n \times n$  skew-symmetric matrix  $A$  with  $n$  odd, show that  $\det(A) = 0$ .

(c) Prove that the trace of any skew-symmetric matrix  $A$  is zero.

**2.E.56.** (a) Find  $A^*$  if  $A = \begin{pmatrix} 1+i & \sqrt{2} & 2-3i \\ 4i & 1-2i & 7 \end{pmatrix}$ . Confirm your answer by Sage.

(b) This is a theoretical task: Show that

- $(zA)^* = \bar{z}A^*$ , for any  $A \in \text{Mat}_{m,n}(\mathbb{C})$  and  $z \in \mathbb{C}$ .
- $(A+B)^* = A^* + B^*$ , for any  $A, B \in \text{Mat}_{m,n}(\mathbb{C})$ .
- $(AB)^* = B^*A^*$ , for any  $A \in \text{Mat}_{m,n}(\mathbb{C})$  and  $B \in \text{Mat}_{n,r}(\mathbb{C})$ .
- $(A^*)^* = A$ , for any  $A \in \text{Mat}_{m,n}(\mathbb{C})$ .
- If  $A \in \text{Mat}_{m,n}(\mathbb{C})$  is invertible, then  $A^*$  is invertible and  $(A^*)^{-1} = (A^{-1})^*$ .
- $\text{tr}(A^*) = \text{tr}(A)$ , for any  $A \in \text{Mat}_m(\mathbb{K})$ , where  $\mathbb{K}$  is one of  $\mathbb{R}$  or  $\mathbb{C}$ . In particular, any  $A \in \text{Mat}_m(\mathbb{R})$  satisfies  $\text{tr}(A^T) = \text{tr}(A)$ .

**Solution.** (a)  $A^*$  is an element of  $\text{Mat}_{3,2}(\mathbb{C})$ , given by  $\begin{pmatrix} 1-i & -4i \\ \sqrt{2} & 1+2i \\ 2+3i & 7 \end{pmatrix}$ . In Sage to compute  $A^*$  use the following cell:

```
A=matrix(CDF, [[1+i, sqrt(2), 2-3*i], [4*i, 1-2*i, 7]])
A_Herm=(A.conjugate()).T
show(A_Herm); bool(A_Herm==A.H)
```

Notice that `A.H` is a shortcut that Sage provides to derive directly the complex conjugate of a complex matrix, analogous to the command `A.T` that we use for the transpose of a matrix (cf. ??). The last line in the previous block verifies that our definition of  $\bar{A}^T$  that we introduced in Sage is consistent with the `A.H` command.

(b) Let  $A \in \text{Mat}_{m,n}(\mathbb{C})$ , with  $A = (a_{ij})$  and  $1 \leq i \leq m$  and  $1 \leq j \leq n$ , respectively. If  $A^* = (a_{k\ell}^*)$ , then  $a_{k\ell}^* = \bar{a}_{\ell k}$  with  $1 \leq k \leq n$  and  $1 \leq \ell \leq m$ , respectively. Consider some  $z \in \mathbb{C}$  and suppose that  $(zA)^* = ((za)_{k\ell}^*)$  and  $\bar{z}A^* = (\bar{z}a_{k\ell}^*)$ , respectively. Then  $(za)_{k\ell}^* = \overline{z\bar{a}_{\ell k}} = \bar{z}\bar{a}_{\ell k} = \bar{z}a_{k\ell}^*$ , which proves that  $(zA)^* = \bar{z}A^*$ . Directly

$$(zA)^* = \overline{zA^T} = (\bar{z}\bar{A})^T = \bar{z}\bar{A}^T = \bar{z}A^*.$$

Next, if  $(A+B)^* = ((a+b)_{k\ell}^*)$ , then  $(a+b)_{k\ell}^* = \overline{(a+b)_{\ell k}} = \bar{a}_{\ell k} + \bar{b}_{\ell k} = a_{k\ell}^* + b_{k\ell}^*$ . This proves the second relation. For the third one, first notice that since  $AB$  is a  $m \times r$  matrix,  $(AB)^*$  is a  $r \times m$  matrix, as the matrix  $B^*A^*$  in the r.h.s. Thus, for any  $k, \ell$  with  $1 \leq k \leq r$  and  $1 \leq \ell \leq m$  we get

$$(ab)_{k\ell}^* = \overline{(ab)_{\ell k}} = \sum_s \overline{a_{\ell s} b_{sk}} = \sum_s \bar{a}_{\ell s} \bar{b}_{sk} = \sum_s \bar{b}_{sk} \bar{a}_{\ell s} = \sum_s b_{ks}^* a_{s\ell}^* = (ba)_{k\ell}^*.$$

Directly  $(AB)^* = \overline{(AB)^T} = \overline{(A\bar{B})^T} = \overline{B^T \bar{A}^T} = B^* A^*$ . For the fourth relation, let  $(A^*)^* = ((a^*)_{k\ell}^*)$  for some  $1 \leq k \leq m$  and  $1 \leq \ell \leq n$ . Then we have  $(a^*)_{k\ell}^* = \overline{a_{k\ell}^*} = \overline{\bar{a}_{\ell k}} = a_{k\ell}$ . Thus  $(A^*)^* = A$ . Now, an application of the conjugate transpose to the identity  $AA^{-1} = E = A^{-1}A$  gives  $(A^{-1})^* A^* = E^* = E = A^*(A^{-1})^*$  and the claim for the inverse follows ( $E$  is

the identity  $n \times n$  matrix).

For the final claim first observe that

$$\operatorname{tr}(A^*) = \sum_{i=1}^m a_{ii}^* = \sum_{i=1}^m \overline{a_{ii}} = \overline{\sum_{i=1}^m a_{ii}} = \overline{\operatorname{tr}(A)}.$$

The result for the real case now follows. □

#### D) MATERIAL ON THE PROPERTIES OF LINEAR MAPPINGS

**2.E.57.** Verify the results of eigenvalues and eigenvectors presented in the following table.

linear transformation	matrix form	eigenvalues	algebraic multiplicity	eigenvectors
identity	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	1	2	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, (V_1 \cong \mathbb{R}^2)$
homothety	$\begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix}, c \in \mathbb{R} \setminus \{0, 1\}$	$c$	2	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, (V_c \cong \mathbb{R}^2)$
rotation	$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}, \theta \neq k\pi, k \in \mathbb{Z}$	$\cos(\theta) \pm i \sin(\theta) = e^{\pm i\theta}$	1, 1	$\begin{pmatrix} 1 \\ i \end{pmatrix}, \begin{pmatrix} 1 \\ -i \end{pmatrix}$
stretching	$\begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix}, c \in \mathbb{R} \setminus \{1\}$	$c, 1$	1, 1	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
hor. shear	$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}, a \in \mathbb{R} \setminus \{0\}$	1	2	$\begin{pmatrix} x \\ 0 \end{pmatrix}, x \in \mathbb{R}$
ver. shear	$\begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}, a \in \mathbb{R} \setminus \{0\}$	1	2	$\begin{pmatrix} 0 \\ y \end{pmatrix}, y \in \mathbb{R}$
reflection (about $y$ -axis)	$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$	$\pm 1$	1, 1	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
reflection (about $x$ -axis)	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	$\pm 1$	1, 1	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

**Solution.** We have explored these transformations and their matrices in detail in 1.E.17, but we will now leave the detailed calculations as an exercise. Note that among the linear transformations in this table, only the vertical or horizontal shears are *not* diagonalizable. The rotation  $R(\theta)$  is diagonalizable over  $\mathbb{C}$  for any  $\theta$ , but it is diagonalizable over  $\mathbb{R}$  only when  $\theta$  is a multiple of  $\pi$ , which we have excluded. □

Consider a  $n \times n$  matrix  $A$  with characteristic polynomial  $\chi_A(\lambda) = \det(A - \lambda E)$ . As described in 2.4.2, this polynomial  $r$  is of degree  $n$ , and can be expressed as

$$\chi_A(\lambda) = c_n \lambda^n + c_{n-1} \lambda^{n-1} + \cdots + c_1 \lambda + c_0,$$

where  $c_n = (-1)^n$ ,  $c_0 = \det(A)$ , and  $c_{n-1} = (-1)^{n-1} \operatorname{tr}(A)$ . For a  $2 \times 2$  matrix this expression simplifies to

$$\chi_A(\lambda) = \lambda^2 - \operatorname{tr}(A) \lambda + \det(A).$$

On the other hand, for a  $3 \times 3$  matrix  $A$  the polynomial reduces to

$$\chi_A(\lambda) = -\lambda^3 + \operatorname{tr}(A) \lambda^2 + c_1 \lambda + \det(A). \quad (*)$$

To use this expression we first need to determine  $c_1$ . An expression of  $c_1$  is provided below.

**2.E.58.** For a  $3 \times 3$  matrix  $A$ , use the relation  $(*)$  to prove that

$$\chi_A(\lambda) = -\lambda^3 + \operatorname{tr}(A) \cdot \lambda^2 + (\det(A - E) + 1 - \operatorname{tr}(A) - \det(A)) \cdot \lambda + \det(A). \quad {}^{14}$$

Next, based on this relation compute the characteristic polynomial and the eigenvalues of the matrix

$$A = \begin{pmatrix} 32 & -67 & 47 \\ 7 & -14 & 13 \\ -7 & 15 & -6 \end{pmatrix}.$$

Can you verify your result in Sage? ○

**2.E.59.** Determine the  $4 \times 4$  real matrix  $A$  whose eigenvalues are given by  $\lambda_1 = 6$  and  $\lambda_2 = 7$  and satisfy one the following conditions:

<sup>14</sup>This method is not the fastest way to compute  $\chi_A$  for a  $3 \times 3$  matrix  $A$ . However, it offers an alternative approach that can be conveniently used to verify our direct computation.

- i)  $\lambda_2$  has algebraic multiplicity three, and geometric multiplicity three;
- ii)  $\lambda_2$  has algebraic multiplicity three, and geometric multiplicity two;
- iii)  $\lambda_2$  has algebraic multiplicity three, and geometric multiplicity one.

**2.E.60.** Find the eigenvalues and the eigenvectors of the matrix  $A = \begin{pmatrix} -1 & -\frac{5}{6} & \frac{5}{3} \\ 0 & -\frac{2}{3} & -\frac{2}{3} \\ 0 & \frac{1}{6} & -\frac{4}{3} \end{pmatrix}$ .

**2.E.61.** In this exercise let  $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$  and consider an idempotent matrix  $A \in \text{Mat}_n(\mathbb{K})$ , i.e.,  $A^2 = A$ . Prove the following statements (see also 2.3.19).

- (a)  $E - A$  is also idempotent, where  $E$  is the identity matrix;
- (b)  $\text{Ker}(A) = \text{Im}(E - A)$  and  $\text{Im}(A) = \text{Ker}(E - A)$ ;
- (c)  $\text{Ker}(A) \oplus \text{Im}(A) = \mathbb{K}^n$ ;
- (d) If  $\mathbb{K}^n = V \oplus W$  for some subspaces  $V, W \subset \mathbb{K}^n$ , then there exists a unique idempotent matrix  $A \in \text{Mat}_n(\mathbb{K})$  such that  $V = \text{Ker}(A)$  and  $W = \text{Im}(A)$ ;
- (e) The eigenvalues of  $A$  are 0 and 1;
- (f)  $\det(A) = 0$  and  $\text{tr}(A) = \text{rank}(A)$ .

**Solution.** (a) We have

$$(E - A)^2 = E - 2A + A^2 = E - 2A + A = E - A.$$

This proves (a).

(b) Suppose that  $y \in \text{Ker}(A)$ , that is,  $Ay = 0$ . Then  $(E - A)y = Ey = y$ , and so  $y \in \text{Im}(E - A)$ . Conversely, if  $y \in \text{Im}(E - A)$ , then we have  $y = (E - A)x$  for some  $x \in \mathbb{K}^n$ . Therefore,  $Ay = A(E - A)x = (A - A^2)x = 0$ , i.e.,  $y \in \text{Ker}(A)$ . This proves that  $\text{Ker}(A) = \text{Im}(E - A)$  and similarly is treated the second relation in (b).

(c) Any  $x \in \mathbb{K}^n$  can be expressed as

$$x = (E - A)x + Ax = v + w, \quad \text{where } v := (E - A)x, \quad w := Ax,$$

respectively. Obviously,  $w \in \text{Im}(A)$  and by (b) we have  $v \in \text{Ker}(A)$ . This shows that  $\mathbb{K}^n = \text{Ker}(A) + \text{Im}(A)$ . Let us now prove that this sum is direct (see 2.3.6). Consider some  $u \in \text{Ker}(A) \cap \text{Im}(A)$ . Because  $u \in \text{Ker}(A)$  means that  $Au = 0$ , and by (b) the relation  $u \in \text{Im}(A)$  implies that  $(E - A)u = 0$ , we get  $u = 0$ , and hence  $\text{Ker}(A) \cap \text{Im}(A) = \{0\}$ . Thus  $\mathbb{K}^n = \text{Ker}(A) \oplus \text{Im}(A)$ .

(d) By assumption  $\mathbb{K}^n = V \oplus W$ , hence any  $x$  is uniquely expressed as  $x = v + w$ , with  $v \in V$  and  $w \in W$ . Consider the mapping  $F : \mathbb{K}^n \rightarrow \mathbb{K}^n$  with  $F(x) = w \in W$ . Clearly,  $F$  is linear and satisfies  $F \circ F = F^2 = F$ . Moreover,  $\text{Im}(F) = W$  and  $\text{Ker}(F) = V$ . Hence, taking as  $A$  the matrix presentation of  $F$  in the standard basis of  $\mathbb{K}^n$ , we are done (the uniqueness follows from the definition of  $F$ ).

(e) Suppose that  $\lambda$  is an eigenvalue of  $A$  with eigenvector  $x$ , i.e.,  $Ax = \lambda x$ . Then

$$A^2x = A(Ax) = A(\lambda x) = \lambda(Ax) = \lambda^2 x,$$

and this shows that  $\lambda^2$  is an eigenvalue of  $A^2$ . However,  $A$  is idempotent and thus we get the relation  $\lambda x = Ax = A^2x = \lambda^2 x$ , that is,  $\lambda(\lambda - 1)x = 0$ . The claim now follows.

(f) Recall that  $\det(A)$  coincides with product of the eigenvalues of  $A$ , hence we get  $\det(A) = 0$  since 0 is an eigenvalue of  $A$ . On the other hand, all eigenvectors of  $A$  belong either to the subspace  $\text{Im}(A)$  (for  $\lambda = 1$ ) or to the subspace  $\text{Ker}(A)$  (for  $\lambda = 0$ ). But since by (c) we have  $\mathbb{K}^n = \text{Ker}(A) \oplus \text{Im}(A)$ , the eigenvectors of  $A$  must span  $\mathbb{K}^n$  and hence  $A$  is diagonalizable. It follows that the multiplicity of  $\lambda = 1$  is precisely the rank of  $A$ . Therefore, the relation  $\text{tr}(A) = \text{rank}(A)$  is now immediate, since the trace coincides with the sum of the eigenvalues  $A$  and the only eigenvalues are 0, 1.

In Chapter 3 we will meet further examples of matrices  $A$  satisfying the decomposition  $\mathbb{K}^n = \text{Ker}(A) \oplus \text{Im}(A)$ , as Hermitian matrices. □

**2.E.62. Eigenface for facial recognition – An application via Sage.** In facial recognition, a prominent technique is the so called *eigenface method*. This technique involves analyzing a set of grayscale images, each represented as a column vector, to identify the *principal component*, which is the eigenvector associated with the largest eigenvalue, also known as the *eigenface*. This eigenface captures the most significant variance among the images.

For an illustration, consider two grayscale images, each with pixel values ranging from 0 to 255, represented by the following  $2 \times 2$  matrices:

$$I_1 = \begin{pmatrix} 50 & 60 \\ 55 & 58 \end{pmatrix}, \quad I_2 = \begin{pmatrix} 70 & 80 \\ 72 & 78 \end{pmatrix}.$$

Using Sage, compute the eigenface by applying the following steps:

- i) Convert the grayscale image into column vectors  $I_1, I_2$  (this is known as “flattening of the image”).
- ii) Construct the data matrix  $D$ : Stack the column vectors  $I_1, I_2$  to create a data matrix.
- iii) Compute the mean image:  $(I_1 + I_2)/2$ .
- iv) Center the data: Subtract the mean image from each column of the data matrix to center the data. Denote this matrix by  $\text{cen}(D)$ .
- v) Calculate the covariance matrix from the centered data:  $\text{cov}_{\text{matrix}} = (\text{cen}(D))^T \text{cen}(D)$ .

**Solution.** Here is a possible solution:

```
# Given images - vectorize the values
I1 = Matrix([[50], [60], [55], [58]])
I2 = Matrix([[70], [80], [72], [78]])
# Construct the data matrix
D = I1.augment(I2)
# Compute the mean image
mean_image = (I1 + I2) / 2
# Subtract the mean from each image in the data matrix
centered_D = D - mean_image * Matrix([1, 1])
# Compute the covariance matrix
cov_matrix = centered_D.transpose() * centered_D
# Find the eigenvalues and eigenvectors
eigenvalues, eigenvectors = cov_matrix.eigenmatrix_right()
# The Eigenface
eigenface_vector = eigenvectors.column(0).normalized()
show(eigenface_vector)
```

Note that the principal component is presented as the primary **Eigenface**, and we have normalized it. Sage’s answer is  $(\frac{1}{2}\sqrt{2}, -\frac{1}{2}\sqrt{2})$ . This is the eigenvector corresponding to the largest eigenvalue of the covariance matrix. It represents the direction of maximum variance in the data, or in other words, the direction along which the variation among the images is the greatest.<sup>15</sup>  $\square$

**2.E.63.** Based on the eigenvalues and eigenvectors of a given  $3 \times 3$  matrix  $A$ , we can often interpret the induced linear transformations on  $\mathbb{R}^3$  geometrically. With this in mind, prove that the following statements are true:

- i) If the matrix has the eigenvalues 0 and 1, the latter with geometric multiplicity 2, then it is a projection in the direction of the eigenvector associated with the eigenvalue 0 onto the plane given by the eigenspace of the eigenvalue 1.
- ii) If the matrix has the eigenvalues 0 and 1, the latter with geometric multiplicity 2, and the eigenvector associated with 0 is perpendicular to the plane generated by the eigenvectors associated with the eigenvalue 1, then the mapping is an orthogonal projection.
- iii) If the matrix has eigenvalue  $-1$  with the eigenvector perpendicular to the plane generated by the eigenvectors associated with the eigenvalue 1, then it is a mirror symmetry through the plane of the eigenvectors associated with 1.
- iv) If the matrix has eigenvalue 1 with an eigenvector perpendicular to the plane generated by the eigenvectors associated with the eigenvalue  $-1$ , then it is an axial symmetry in  $\mathbb{R}^3$  through the axis given by the eigenvector associated with 1.

**Solution.** We will prove the first three statements and leave the last one as an exercise.

(i) Since 1 has algebraic multiplicity 2 there are two linearly independent vectors  $\{v_1, v_2\}$  corresponding to this eigenvalue. Since the algebraic multiplicity of 0 is 1, the geometric multiplicity of 0 should be also 1. This means that the corresponding eigenspace (nullspace)  $V_0$  generated by a unique vector  $v_0$ , and we get a direct sum decomposition  $\mathbb{R}^3 = V_0 \oplus V_1$ , with  $\dim V_0 = 1$  and  $\dim V_1 = 2$ , respectively. Thus, any vector on  $u \in \mathbb{R}^3$  can be expressed as  $u = \sum_{i=0}^2 c_i v_i$ , for some scalars  $c_0, c_1, c_2$ . By applying the  $3 \times 3$  matrix  $A$  to  $x$  we get

$$Au = A(c_0 v_0 + c_1 v_1 + c_2 v_2) = c_0 A v_0 + c_1 A v_1 + c_2 A v_2 = c_1 v_1 + c_2 v_2.$$

It is now easy to prove that  $A^2 = A$ :

$$A^2(u) = A(Au) = A(c_1 v_1 + c_2 v_2) = c_1 A v_1 + c_2 A v_2 = c_1 v_1 + c_2 v_2 = Au, \quad u \in \mathbb{R}^3.$$

<sup>15</sup>Principal Component Analysis (PCA) is a statistical method used to identify patterns in data and reduce dimensionality by projecting the data onto a set of orthogonal axes. In the context of PCA and the eigenface method, the *maximum variance* in the data is the direction in which the data varies the most.



Thus,  $A$  is a projection matrix mapping any vector  $u \in \mathbb{R}^3$  onto the plane  $V_1$  along the direction defined by the line  $V_0$  in  $\mathbb{R}^3$ .  
 (ii) Let us use the notation from (i) and assume that the eigenvector  $v_0$  associated with 0 is perpendicular to  $V_1$ . Without loss of generality, we may assume that the eigenvectors  $v_1, v_2$  that span  $V_1$  are orthonormal, i.e.,

$$\langle v_1, v_1 \rangle = v_1^T v_1 = 1 = v_2^T v_2 = \langle v_2, v_2 \rangle, \quad \langle v_1, v_2 \rangle = v_2^T v_1 = 0 = v_1^T v_2 = \langle v_2, v_1 \rangle.$$

Since  $v_1 v_1^T$  and  $v_2 v_2^T$  are the projection matrices onto the lines spanned by  $v_1, v_2$ , respectively, the projection matrix  $A$  onto the plane  $V_1$  has the form  $A = v_1 v_1^T + v_2 v_2^T$ . It is now a direct computation to prove that  $A$  is an orthogonal projection, i.e.,  $A^2 = A$  and  $A = A^T$ .

(iii) By assumption, the matrix  $A$  has the eigenvalues  $-1$  and  $1$ , the second one with geometric multiplicity 2. Thus, there are two linearly independent vectors  $u_1, u_2$  which generate the eigenspace  $V_1$ . Consequently, any vector  $x \in V_1$  can be written as a linear combination of  $u_1, u_2$ . Also,  $V_{-1}$  is 1-dimensional, generated by the eigenvector corresponding to the eigenvalue  $-1$ , which we may denote by  $v_0$ . This eigenvector is perpendicular to  $V_1$  and hence we have  $\langle v_1, v_0 \rangle = \langle v_2, v_0 \rangle = 0$ . Let  $x = a_1 u_1 + a_2 u_2$  be some arbitrary vector of  $V_1$ . Then we see that

$$Ax = A(a_1 u_1 + a_2 u_2) = a_1 A u_1 + a_2 A u_2 = a_1 u_1 + a_2 u_2 = x,$$

i.e.,  $A$  preserves  $V_1$ . Moreover, since  $A v_0 = -v_0$  we deduce that  $v_0$  is reflected through the plane  $V_1$ . Since  $A$  reflects vectors that are perpendicular to the plane  $V_1$ , while leaving vectors in  $V_1$  unchanged, we deduce that  $A$  represents a mirror symmetry through the plane  $V_1$  in  $\mathbb{R}^3$ .  $\square$

**2.E.64.** (a) Determine the eigenvalues and eigenvectors of the matrix

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix}.$$

(b) Describe the geometric interpretation of the corresponding linear mapping, and write down its matrix in the basis  $\underline{u} = \{u_1 = (1, -1, 1)^T, u_2 = (1, 2, 0)^T, u_3 = (0, 1, 1)^T\}$ .

**Solution.** (a) The characteristic polynomial  $\chi_A(\lambda) = \det(A - \lambda E)$  of  $A$  is of degree three, so over  $\mathbb{C}$  we expect to find three roots. We compute

$$\begin{vmatrix} 1-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 1 & 2 & 1-\lambda \end{vmatrix} = -\lambda^3 + 4\lambda^2 - 2\lambda = -\lambda(\lambda^2 - 4\lambda + 2)$$

and hence the equation  ${}_A P(\lambda) = 0$  has only real solutions, given by  $\lambda_1 = 0$ ,  $\lambda_{\pm} = 2 \pm \sqrt{2}$ . These are the three eigenvalues of  $A$ . To verify these claims we can use Sage and apply the rules from 2.D.2. Thus, let us type

```
A=matrix(SR, [[1, 1, 0], [1, 2, 1], [1, 2, 1]])
show(A)
p(x) = A.characteristic_polynomial(x)
show(p(x)); show(p.roots())
```

Sage prints out the characteristic polynomial of  $A$  (up to a sign), and the triple

$$\left[ \left( -\sqrt{2} + 2, 1 \right), \left( \sqrt{2} + 2, 1 \right), (0, 1) \right],$$

where we can read the three eigenvalues together with the multiplicities. Alternatively, we can use the command `A.eigenvalues()`. Next we should compute the corresponding eigenvectors:

• For  $\lambda_1 = 0$  we get the kernel of  $A$ , which is of course determined by solving the matrix equation  $Ax = 0$ , i.e.,

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The result is  $V_0 = \text{Ker}(A) = \text{span}_{\mathbb{R}}\{E_0 := (1, -1, 1)^T\}$ , and hence the eigenspace  $V_0$  is 1-dimensional.

• For  $\lambda_+ = 2 + \sqrt{2}$  we get the matrix equation

$$\begin{pmatrix} -(1 + \sqrt{2}) & 1 & 0 \\ 1 & -\sqrt{2} & 1 \\ 1 & 2 & -(1 + \sqrt{2}) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Again the solution forms a one-dimensional space,  $V_{\lambda_+} = \text{span}_{\mathbb{R}}\{E_+ := (1, 1 + \sqrt{2}, 1 + \sqrt{2})^T\}$ .

• For  $\lambda_- = 2 - \sqrt{2}$  we solve the matrix equation

$$\begin{pmatrix} (\sqrt{2}-1) & 1 & 0 \\ 1 & \sqrt{2} & 1 \\ 1 & 2 & (\sqrt{2}-1) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Solving the corresponding homogeneous linear system we see that  $V_{\lambda_-} = \text{span}_{\mathbb{R}}\{E_- := (1, 1 - \sqrt{2}, 1 - \sqrt{2})^T\}$ . To verify the results for the eigenvectors using Sage, just add the code `A.eigenvectors_right()`. Of course, the case of the zero-eigenvalue which corresponds to the null space of  $A$  can be treated by the command `A.right_kernel()`, as well.

(b) Let us first express the induced linear transformation in the given basis  $\underline{u}$ , which is easier and relies on  $A$ . For this, we need to determine the matrix  $T$  for moving from the standard basis  $\underline{e}$  to the new basis  $\underline{u}$ . But let us apply the tip from 2.C.41, and first obtain the matrix for changing the basis from  $\underline{u}$  to the standard one, that is, the matrix  $T^{-1}$ , whose columns consist just of the vectors  $u_1, u_2, u_3$ :

$$T^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix} \Rightarrow T = (T^{-1})^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ -2 & 1 & 3 \end{pmatrix}.$$

Hence, the matrix of the linear transformation under the new basis is given by

$$TAT^{-1} = \begin{pmatrix} 0 & 5 & 2 \\ 0 & -2 & -1 \\ 0 & 14 & 6 \end{pmatrix}.$$

To determine geometrically the mapping induced by the given matrix  $A$ , we can combine our results about the eigenvectors of  $A$  with the statements in 2.E.63. The induced map can be interpreted as a projection along the eigenvector  $E_0 \in V_0 = \text{Ker}(A)$  into the plane generated by the eigenvectors  $E_+ \in V_+$  and  $E_- \in V_-$ , composed with the linear mapping given by stretching in the directions of these eigenvectors by the factor corresponding to the eigenvalues.  $\square$

**2.E.65.** Given the matrix

$$M = \begin{pmatrix} -\frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{4}{3} & -\frac{7}{3} & -\frac{8}{3} \\ -1 & 1 & -1 \end{pmatrix}$$

determine geometrically the induced linear mapping (in a similar way we did in task 2.E.64).  $\circ$

**2.E.66. (Pauli matrices)** In physics, the state of a particle with spin  $\frac{1}{2}$  is described by the so called *Pauli matrices*. They are the following  $2 \times 2$  matrices over  $\mathbb{C}$ :

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

where as usual  $i = \sqrt{-1}$ . For square matrices we define their *commutator* (denoted by square brackets  $[ , ]$ ) as  $[\sigma_1, \sigma_2] := \sigma_1\sigma_2 - \sigma_2\sigma_1$ .

i) Show that  $[\sigma_1, \sigma_2] = 2i\sigma_3$ ,  $[\sigma_1, \sigma_3] = -2i\sigma_2$ , and  $[\sigma_2, \sigma_3] = 2i\sigma_1$ , respectively.

ii) Prove that  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$  and that the eigenvalues of the matrices  $\sigma_1, \sigma_2, \sigma_3$  are  $\pm 1$ .

iii) It is known that the matrices describing the state of the particle with spin 1 are the following  $3 \times 3$  matrices over  $\mathbb{C}$ :

$$\sigma_x = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \sigma_y = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -i & 0 \\ i & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Show that these matrices satisfy the same commuting relations as the Pauli matrices.  $\circ$

**2.E.67. (Quaternionic relations)** Using the Pauli matrices  $\sigma_1, \sigma_2, \sigma_3$  introduced in 2.E.66, set

$$I := E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad J := i\sigma_3, \quad K := i\sigma_2, \quad L := i\sigma_1.$$

Prove the following relations, known as the *quaternionic relations*:

$$I^2 = J^2 = K^2 = -1, \quad IJ = -JI = K, \quad JK = -KJ = I, \quad KI = -IK = J.$$

**2.E.68. Remark.** The quaternionic relations impose on the 4-dimensional space spanned by  $1, I, J, K$  the structure of a 4-dimensional *algebra* over  $\mathbb{R}$ . This algebra was invented by W. R. Hamilton in 1843, and is known as the *algebra of quaternions*, denoted by  $\mathbb{H}$ , that is,

$$\mathbb{H} = \text{span}_{\mathbb{R}}\{1, I, J, K\}.$$

Roughly speaking, an “algebra” is a vector space endowed with a binary bilinear operation for multiplication. For the algebra of quaternions, the multiplication is given by the matrix multiplication, hence observe that  $\mathbb{H}$  is associative and non-commutative. A *quaternion* is simply an element in  $\mathbb{H}$ , and hence a linear combination of  $\{1, I, J, K\}$ . On the other hand, the algebra generated by the matrices  $\sigma_1, \sigma_2, \sigma_3$  is isomorphic (as algebra) to the so called *Clifford algebra* of  $\mathbb{R}^3$ . This Clifford algebra is strongly related with developments in geometric algebra, which is an advanced topic with beautiful applications in robotics.

## Solutions to the problems

2.A.2. 1) We have  $(A-E)(A+E) = A^2 + AE - EA - E^2 = A^2 - E$ , since  $AE = EA = A$  for any square matrix  $A$  and the identity matrix satisfies  $E^n = E$  for all  $n$ . We compute  $A^2 = \begin{pmatrix} 1 & 1 + \sqrt{2} \\ 0 & 2 \end{pmatrix}$ , and hence  $(A-E)(A+E) = \begin{pmatrix} 0 & 1 + \sqrt{2} \\ 0 & 1 \end{pmatrix}$ .

2) Here we get  $AB = \begin{pmatrix} 5 & 4 \\ 4\sqrt{2} & 0 \end{pmatrix}$ , thus  $A(Bu) = (AB)u = \begin{pmatrix} 5 & 4 \\ 4\sqrt{2} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix} = \begin{pmatrix} 4\sqrt{2} \\ 0 \end{pmatrix}$ . Alternatively, we could first compute  $Bu$ , that is,

$$Bu = \begin{pmatrix} 1 & 4 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix} = \begin{pmatrix} 4\sqrt{2} \\ 0 \end{pmatrix} \implies A(Bu) = \begin{pmatrix} 1 & 1 \\ 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 4\sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 4\sqrt{2} \\ 0 \end{pmatrix}.$$

3) In this case we compute

$$(A-B)(A+B) = \begin{pmatrix} 0 & -3 \\ -4 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 5 \\ 4 & \sqrt{2} \end{pmatrix} = \begin{pmatrix} -12 & -3\sqrt{2} \\ 4\sqrt{2}-8 & -18 \end{pmatrix}.$$

Thus  $(A-B)(A+B)u = \begin{pmatrix} -12 & -3\sqrt{2} \\ 4\sqrt{2}-8 & -18 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix} = -\begin{pmatrix} 6 \\ 18\sqrt{2} \end{pmatrix} = -6 \begin{pmatrix} 1 \\ 3\sqrt{2} \end{pmatrix}$ .

4) Obviously

$$aA^2 + aB^2 = a(A^2 + B^2) = a \begin{pmatrix} 18 & \sqrt{2} + 5 \\ 4 & 18 \end{pmatrix}.$$

5) The inner product of  $u, v$  is given by

$$u \cdot v = \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 4 \end{pmatrix} = 0 \cdot 1 + \sqrt{2} \cdot 4 = 4\sqrt{2},$$

hence the given quantity equals to zero.

6) This is also easy. We compute  $A^3 = AA^2 = \begin{pmatrix} 1 & 1 \\ 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 1 & 1 + \sqrt{2} \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 3 + \sqrt{2} \\ 0 & 2\sqrt{2} \end{pmatrix}$ , and thus

$$4A^3 - B^2 = \begin{pmatrix} 4 & 12 + 4\sqrt{2} \\ 0 & 8\sqrt{2} \end{pmatrix} - \begin{pmatrix} 17 & 4 \\ 4 & 16 \end{pmatrix} = \begin{pmatrix} -13 & 8 + 4\sqrt{2} \\ -4 & -16 + 8\sqrt{2} \end{pmatrix}.$$

7) In this case we obtain  $\sqrt{2}u - \frac{1}{2}v = \begin{pmatrix} 0 \\ 2 \end{pmatrix} - \begin{pmatrix} 1/2 \\ 2 \end{pmatrix} = \begin{pmatrix} -1/2 \\ 0 \end{pmatrix}$ .

8) Above we computed  $A(Bu) = \begin{pmatrix} 4\sqrt{2} \\ 0 \end{pmatrix}$ . Hence

$$\frac{\sqrt{2}}{2}u - ABu + 2v = \begin{pmatrix} 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 4\sqrt{2} \\ 0 \end{pmatrix} + \begin{pmatrix} 2 \\ 8 \end{pmatrix} = \begin{pmatrix} 2 - 4\sqrt{2} \\ 9 \end{pmatrix}.$$

9) In this case we have

$$A^2u - B^2v = \begin{pmatrix} 1 & 1 + \sqrt{2} \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix} - \begin{pmatrix} 17 & 4 \\ 4 & 16 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 + \sqrt{2} \\ 2\sqrt{2} \end{pmatrix} - \begin{pmatrix} 33 \\ 68 \end{pmatrix} = \begin{pmatrix} \sqrt{2} - 31 \\ 2\sqrt{2} - 68 \end{pmatrix}.$$

10) Notice that  $B^2u = \begin{pmatrix} 17 & 4 \\ 4 & 16 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{2} \end{pmatrix} = \begin{pmatrix} 4\sqrt{2} \\ 16\sqrt{2} \end{pmatrix} = 4\sqrt{2}v$ . Hence the vector  $w := B^2u$  is a non-zero multiple of the vector  $v$ , and since  $\|v\|^2 = v \cdot v$ , we see that the given expression is a multiple of the square of the norm of  $v$ . Indeed, we have  $v \cdot v = 17$  and a direct computation gives

$$(B^2u) \cdot (4v) = w \cdot (4v) = (4\sqrt{2}v) \cdot (4v) = 16\sqrt{2}v \cdot v = 16\sqrt{2}\|v\|^2 = 272\sqrt{2}.$$

2.A.4. Notice that one could follow alternative computations to obtain the same results. Now, for a verification of the results in Sage, we use the cell

```
A=matrix([[1, sqrt(2), pi], [0, 1, 1], [-pi, 0, 2]])
B=matrix([[0, 1, 1], [-1, 0, pi], [-1, 0, 1]])
u=vector([pi, 0, sqrt(2)]); E=matrix.identity(3)
show(A-E); show((A-E)^2)
C=(A-E)^2*(4*B); show(C); show(C*u)
show(B^2); show((A-E).transpose())
D=(A-E).transpose()*B^2; show(D); show(D*u)
show(A*B*u); show(B*A*u)
```

**2.A.16.** (a) The linear systems whose extended matrix is  $B_{\pm}$ , are overdetermined, i.e., we have more equations than unknowns:

$$\underbrace{\begin{pmatrix} 1 & 1 \\ 3 & 2 \\ 2 & \pm 1 \end{pmatrix}}_{3 \times 2} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{2 \times 1} = \underbrace{\begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix}}_{3 \times 1},$$

or in other words

$$\begin{cases} x_1 + x_2 = -2, \\ 3x_1 + 2x_2 = 0, \\ 2x_1 + x_2 = 2, \end{cases} \quad \text{and} \quad \begin{cases} x_1 + x_2 = -2, \\ 3x_1 + 2x_2 = 0, \\ 2x_1 - x_2 = 2, \end{cases}$$

respectively. Such systems are inconsistent, unless some of the equations are linearly dependent. For the matrix  $B_+$ , the cell

```
Bp=matrix(QQ, [[1, 1, -2], [3, 2, 0], [2, 1, 2]])
rec=Bp.rref();show(rec)
print(Bp.pivots())
```

returns the matrix  $\begin{pmatrix} 1 & 0 & 4 \\ 0 & 1 & -6 \\ 0 & 0 & 0 \end{pmatrix}$ , and the pair  $(0, 1)$ , respectively. Thus, the pivots are on the first and second column and the system admits a unique solution, given by  $\{x_1 = 4, x_2 = -6\}$ . This also follows by using the backslash operator:

```
Ap=matrix(QQ, [[1, 1], [3, 2], [2, 1]])
b=vector([-2, 0, 2])
x=Ap\b; show(x)
```

Theoretically, this unique solution is explained by the fact that the second equation in the corresponding linear system is a linear combination of the first and the third:

$$3x_1 + 2x_2 = (x_1 + x_2) + (2x_1 + x_2) = -2 + 2 = 0.$$

On the other side for the matrix  $B_-$  Sage gives as reduced row echelon form the identity  $3 \times 3$  matrix, and so as pivot columns all the three columns, i.e., the pivot command for  $B_-$  returns  $(0, 1, 2)$ . The existence of a pivot in the last column implies that the corresponding linear system is inconsistent, a claim that you can easily prove also by hand.

(b) For the matrix  $C$  the cell

```
C=matrix(QQ, [[2, 1, 7, 0], [-1, 4, 10, 5], [3, 2, 12, 10], [0, 0, 0, 10]])
show(C.rref())
print(C.pivots())
```

returns

$$\begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad (0, 1, 3),$$

respectively. Thus, the pivots are located in the first, second, and fourth columns. Moreover, the presence of the last row containing only zeros alerts the possibility of an infinite number of solutions. Indeed, because there is no pivot in the 3rd column,  $x_3$  is a free variable. In particular, the homogeneous system  $Cx = 0$  has infinite solutions given by  $x_1 = -2x_3 = -2t$ ,  $x_2 = -3x_3 = -3t$ ,  $x_4 = 0$ , with  $x_3 = t \in \mathbb{R}$ .

**2.A.17.** Clearly the unknowns  $X_1$  and  $X_2$  must be  $2 \times 2$  matrices, so we may set  $X_1 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$  and  $X_2 = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$ , respectively, for some  $a_i, b_i, c_i, d_i \in \mathbb{R}$ , for all  $i = 1, 2$ . Then, for the first case an easy computation gives that

$$\begin{pmatrix} a_1 + 3c_1 & b_1 + 3d_1 \\ 3a_1 + 8c_1 & 3b_1 + 8d_1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

This implies the following system of equations  $\{a_1 + 3c_1 = 1, b_1 + 3d_1 = 2, 3a_1 + 8c_1 = 3, 3b_1 + 8d_1 = 4\}$ , which has a unique solution given by  $a_1 = 1, b_1 = -4, c_1 = 0$  and  $d_1 = 2$ . Thus  $X_1 = \begin{pmatrix} 1 & -4 \\ 0 & 2 \end{pmatrix}$ .

Let apply a more elegant method for the second case, which is based on the inverse of a  $2 \times 2$  matrix  $A$ . Recall that the inverse of a  $n \times n$  matrix is the square matrix of the same order as  $A$ , uniquely determined by the equation  $AA^{-1} = A^{-1}A = E$ , where as usual  $E$  is the identity matrix. Notice that  $A^{-1}$  may not always exist. By the discussion in 2.2.11 we know the following nice expression of the inverse of a  $2 \times 2$  matrix  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ , where  $a, b, c, d \in \mathbb{R}$  and  $\det(A) = ad - bc \neq 0$ :

$$A^{-1} = \frac{1}{(ad-bc)} \cdot \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

An application gives  $A^{-1} = \begin{pmatrix} -8 & 3 \\ 3 & -1 \end{pmatrix}$ , and then by multiplying both sides of the given equation from the right by  $A^{-1}$ , we deduce that

$$X_2 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} A^{-1} = \begin{pmatrix} -2 & 1 \\ -12 & 5 \end{pmatrix}$$

(since  $X_2 A A^{-1} = X_2 I = X_2$ ). Of course, this method applies for solving also the first matrix equation (here one should multiply both sides of the given equation by the corresponding  $A^{-1}$  from the left, so that  $A^{-1} A X_1 = I X_1 = X_1$ ).

**2.A.19.** We obtain

$$A^{-1} = \begin{pmatrix} 1 & 10 & -4 \\ 1 & 12 & -5 \\ 0 & 5 & -2 \end{pmatrix}, B^{-1} = \begin{pmatrix} 2 & -3 & 0 & 0 & 0 \\ -5 & 8 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -5 & 2 \\ 0 & 0 & 0 & 3 & -1 \end{pmatrix}, C^{-1} = \frac{1}{2} \cdot \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}, D^{-1} = \frac{1}{2} \cdot \begin{pmatrix} 3 & -i \\ i & 1 \end{pmatrix}.$$

**2.A.20.** Of course, we may use Sage to transform the given matrix  $A$  to a row-echelon form.

```
A=matrix([[1, -3, 0, 1], [1, -2, 2, -4], [1, -1, 0, 1], [-2, -1, 1, -2]])
show(A.echelon_form())
```

Sage returns the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 6 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 10 \end{pmatrix},$$

hence the rank of  $A$  equals to four. In fact, Sage allows us to directly compute the rank of a given matrix  $A$ , by the command `A.rank()`.

Let us move to the remaining questions. The extended matrix of the first linear system has the form

$$\left( \begin{array}{cccc|c} 1 & 1 & 1 & -2 & 4 \\ -3 & -2 & -1 & -1 & 5 \\ 0 & 2 & 0 & 1 & 1 \\ 1 & -4 & 1 & -2 & 3 \end{array} \right).$$

But the left-hand side of this extended matrix is exactly  $A^T$ , thus we can get the column-echelon form the same way as before. In particular, the columns of the matrix are linearly independent and the rank is maximal, i.e., four again. Therefore, there exists the inverse of  $A^T$ , i.e., the matrix  $(A^T)^{-1}$ , and it turns out that the system has a unique solution given by

$$(x_1, x_2, x_3, x_4)^T = (A^T)^{-1} \cdot (4, 5, 1, 3)^T.$$

Next, the homogenous system has the same left-hand side (given by the matrix  $A^T$ ), as the first. Because the numbers on the right-hand side of the equations in the system do not influence the number of solutions and because every homogeneous system has a zero solution, the only solution of the second system is the zero one,  $(x_1, x_2, x_3, x_4)^T = (0, 0, 0, 0)^T$ .

The extended matrix of the 3rd system coincides with  $A$ :

$$\left( \begin{array}{ccc|c} 1 & -3 & 0 & 1 \\ 1 & -2 & 2 & -4 \\ 1 & -1 & 0 & 1 \\ -2 & -1 & 1 & -2 \end{array} \right).$$

Sage revealed to us that the transformation of the matrix into the row echelon form, yields a row of the form

$$(0 \ 0 \ 0 \ | \ a), \quad \text{where } a \neq 0.$$

However, the column on the right-hand side is not a linear combination of the columns on the left-hand side (the rank of the matrix is four). It follows that this system has no solution. We may verify this claim easily in Sage, e.g. via the `solve` command.

```
var("x1, x2, x3")
eq1=x1-3*x2-1;eq2=x1-2*x2+2*x3+4
eq3=x1-x2-1;eq4=-2*x1-x2+x3+2
solve([eq1==0,eq2==0,eq3==0,eq4==0],x1,x2,x3)
```

Sage's output is `[]`, indicating that the system has no solutions (the output `[]` is a common way for Sage to represent an empty solution set).

**2.B.2.** First we treat  $\sigma$ , where observe that after the integer 3 in the second row, there are the small integers 1 and 2, so in the required sum we add 2. Next, after 1 there is no smaller number and we add 0. In the sequel, after the integer 6 there are the smaller numbers five, four and two, thus we add 3, similarly for seven, eight and nine, while for five we add 2, for four we add 1 and for two nothing. Thus we have  $2 + 3 + 3 + 3 + 3 + 2 + 1 = 17$  inversions in total, which implies that  $\sigma$  is an odd permutation. Its parity is  $(-1)^{17} = -1$ . Applying the alternative method (using the theorem given in 2.2.2), we obtain the expression  $\sigma = (1, 3)(3, 6)(6, 9)(9, 2)(4, 7)(7, 5)(5, 8)$ . Thus, there are seven transpositions in the decomposition, and the permutation is indeed odd.

Similarly we can treat the second permutation  $\tau$ , which decomposes into a product of three transpositions (using the cycle decomposition):  $\tau = (1, 2, 4)(3, 6) = (1, 2)(2, 4)(3, 6)$ . Or one can count the number of inversions in  $\tau$ . The latter is given by  $1 + 2 + 3 + 0 + 1 = 7$ , and thus both ways imply that  $\tau$  is an odd permutation, i.e.,  $\text{sign}(\tau) = -1$ .

**2.B.7.** According to Laplace theorem given 2.2.9 we can compute the determinant via expansion along the fifth row (or the second column, which each contains four zeros). We obtain:

$$\begin{vmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 3 & 0 & 3 \\ 4 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 5 \end{vmatrix} = 0 \cdot \begin{vmatrix} 0 & 1 & 0 & 1 \\ 2 & 0 & 2 & 0 \\ 0 & 3 & 0 & 3 \\ 0 & 0 & 4 & 4 \end{vmatrix} - 0 \cdot \begin{vmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 3 & 0 & 3 \\ 4 & 0 & 4 & 4 \end{vmatrix} + 0 \cdot \begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 4 & 0 & 4 & 4 \end{vmatrix} - 0 \cdot \begin{vmatrix} 1 & 0 & 1 & 1 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 3 \\ 4 & 0 & 0 & 4 \end{vmatrix} + 5 \cdot \begin{vmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 3 & 0 \\ 4 & 0 & 0 & 4 \end{vmatrix}$$

$$= 5 \cdot \begin{vmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 3 & 0 \\ 4 & 0 & 0 & 4 \end{vmatrix} = 5 \cdot 2 \cdot \begin{vmatrix} 1 & 1 & 0 \\ 0 & 3 & 0 \\ 4 & 0 & 4 \end{vmatrix} = 120.$$

The determinant of the last  $3 \times 3$  matrix was computed by the Sarrus rule. The check in Sage is most simple:

```
D=matrix([[1, 0, 1, 0, 1],[0, 2, 0, 2, 0],[0, 0, 3, 0, 3],[4, 0, 0, 4, 4],[0, 0, 0, 0, 5]])
det(D)
```

**2.B.8.** (a) This is a direct consequence of the Cauchy theorem, that is,  $\det(AB) = \det(A) \det(B)$ , see 2.2.7.

(b) This identity is in fact a consequence of the following: The determinant of a  $n \times n$  matrix  $A$  is a *multilinear mapping* in the columns (or rows) of  $A$ . Hence for example if  $c \in \mathbb{R}$  and  $A^j$  denotes the  $j$ th column of  $A$  such that  $A = (A^1 \ A^2 \ \dots \ A^n)$ , then we have

$$\det(cA) = \det((cA^1 \ cA^2 \ \dots \ cA^n)) = c^n \det((A^1 \ A^2 \ \dots \ A^n)) = c^n \det(A).$$

It remains to prove our statement for the multi-linearity of  $\det$ , which we leave as an exercise on multilinear mappings, see also 2.3.23.

(c) On can present many examples. For instance, based on the identity presented in part (b) for  $A = B$  and  $n > 1$  we see that  $\det(A + B) = \det(2A) = 2^n \det(A) > 2 \det(A) = \det(A) + \det(B)$ .

**2.B.10.** (a) We present the solutions mainly via Sage, and leave to the reader the description of formal proofs. To find the null space of the matrix  $A$ , give the cell

```
A=matrix([[1, -1, pi, 0],[0, pi, -1, 0], [1, 0, -1, pi]])
show(A.right_kernel())
```

The output has the form

$$\text{RowSpan}_{\text{SR}} \left( 1 \quad -\frac{1}{\pi(\pi-\frac{1}{\pi})} \quad -\frac{1}{\pi-\frac{1}{\pi}} \quad -\frac{\pi-\frac{1}{\pi}+1}{\pi(\pi-\frac{1}{\pi})} \right).$$

Therefore,  $\text{Ker}(A)$  is 1-dimensional. In a similar way we compute  $\text{Ker}(B) = \text{span}\{(1, -1, -1, 1)^T\}$ . We will learn for vector subspaces and the notion of “dimension” of a vector (sub)space very soon, in Section C.

(b) Obviously, the required submatrix has the form  $\begin{pmatrix} \pi & 0 \\ 0 & \pi \end{pmatrix}$ , which is diagonal. In Sage there are many ways to determine submatrices. For the specific case, where we should cancel both rows and columns, an appropriate command is `A.matrix_from_rows_and_columns()`:

```
d=A.matrix_from_rows_and_columns([1, 2],[1,3]); show(d)
#verify that d is diagonal
print(bool(d==diagonal_matrix([pi, pi])))
```

In this command the brackets  $[1, 2]$  and  $[1, 3]$  specify the rows and columns that are retained in the submatrix.

Another method relies on the bracket operation that we shortly mentioned above. Hence for example, add the cell

```
suba=A[1:3, 1:4:2]; show(suba)
```

We mention that in the expression  $A[1 : 3, 1 : 4 : 2]$  the final index 2 is the slice step between the columns.

(c) The submatrix obtained by  $B$  after canceling the first and last row, and the fourth column has the form  $\begin{pmatrix} 2 & 1 & 4 \\ 3 & 4 & 1 \end{pmatrix}$ . In Sage use the block

```
B=matrix([[1, 2, 3, 4],[2, 1, 4, 3], [3, 4, 1, 2], [4, 3, 2, 1]])
show(B.matrix_from_rows_and_columns([1, 2],[0, 1,2]))
```

or, as an alternative, you may type

```
B=matrix([[1, 3, 4],[2, 1, 4, 3], [3, 4, 1, 2], [4, 3, 2, 1]])
show(B[1:3, 0:-1])
```

In fact, one can replace the final line by the command `show(B[1 : 3, 0 : 3])`.

(d) In this case the required submatrix has the form  $\begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$ , which is obviously symmetric. In Sage add in the previous block the code

```
subb=B.matrix_from_rows_and_columns([1,3], [0, 2])
bool(subb==subb.transpose())
```

**2.C.2.** The set of real polynomials of degree exactly  $m$  is *not* a real vector space, because it fails to satisfy one of the key properties of a vector space: closure under addition. This is because the sum of two polynomials of degree  $m$  does not necessarily have the same degree. Can you provide an explicit example?

**2.C.4.** One can use the following block:

```
v1=vector(QQ, [1/2, 2/3, 1, 0])
v2=vector(QQ, [2, 0, 1/7, 0])
v3=vector(QQ, [0, 1/5, 2/5, 1])
v4=vector(QQ, [-1, 2, 0, 3])
show(3*v1+2*v2+5*v3+v4)
```

Another approach involves placing the vectors and their coefficients into two separate lists and then manipulating these lists as follows:

```
vectors=[v1, v2, v3, v4]
scalars=[3, 2, 5, 1]
lin_comb=sum([scalars[i]*vectors[i] for i in range(len(vectors))])
show(lin_comb)
```

In this block we used the `len()` function to determine the number of elements in the list with vectors, which can be very useful when working with large lists. Check yourselves that both cells print out the following answer:  $(\frac{9}{2}, 5, \frac{37}{7}, 8)$ . There are further methods to describe the linear combination (span) of vectors via Sage, presented in [2.C.12](#).

**2.C.5.** A solution goes as follows:

```
A=matrix(QQ, [[3, -9, 9, 0],[9, -3, 6, 0],[9, -6, 0,-6],[-9, 12, 6, 6]])
b = vector(QQ, [-1, -1, 2, 0])
cols=A.columns(); show(cols)
soln = [-2/15, 1/5, 2/15, -11/15]
bool(sum([soln[i]*cols[i] for i in range(len(cols))])==b)
```

In this block we used the function `.columns()` to access the columns of  $A$ . Note also that the final line verifies that linear combination of the columns of  $A$  with coefficients from `soln` (this represents the given solution), equals the vector  $b$ . Executing this block we obtain `True`. In fact, in our solution we could replace the final line by

```
show(sum([soln[i]*cols[i] for i in range(len(cols))]))
```

so that we can directly compare this expression with the constant vector  $b$ .



**2.C.6.** The finite field  $\mathbb{Z}_5$  is an important mathematical structure used in various areas (such as algebra, cryptography, and computer science). It consists of the integers from 0 to 4,  $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$ , with arithmetic operations performed modulo 5. For your convenience we present the addition and multiplication table for  $\mathbb{Z}_5$ :

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

For instance,  $3 + 4 = 7 = 2(\text{mod}5)$  and  $2 \cdot 3 = 6 = 1(\text{mod}5)$ . The vector space  $\mathbb{Z}_5^3 = \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z}_5$  consists of vectors  $(a, b, c)^T$ , where each entry  $a, b, c$  is an element of the finite field  $\mathbb{Z}_5$ . Hence,  $\mathbb{Z}_5^3$  is 3-dimensional over  $\mathbb{Z}_5$ . The addition and scalar multiplication are respectively defined by

$$u + v = \begin{pmatrix} (u_1 + v_1) \text{ mod } 5 \\ (u_2 + v_2) \text{ mod } 5 \\ (u_3 + v_3) \text{ mod } 5 \end{pmatrix}, \quad cu = \begin{pmatrix} (cu_1) \text{ mod } 5 \\ (cu_2) \text{ mod } 5 \\ (cu_3) \text{ mod } 5 \end{pmatrix},$$

for any two vectors  $u = (u_1, u_2, u_3)^T, v = (v_1, v_2, v_3)^T$  and scalar  $c \in \mathbb{Z}_5$ . For instance:

$$\begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} (3+2) \text{ mod } 5 \\ (4+3) \text{ mod } 5 \\ (2+4) \text{ mod } 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}, \quad 3 \begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \text{ mod } 5 \\ 12 \text{ mod } 5 \\ 6 \text{ mod } 5 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}.$$

The zero vector is the additive identity in  $\mathbb{Z}_5^3$ . Can you derive the expression of the additive inverse of some vector  $u = (u_1, u_2, u_3)^T \in \mathbb{Z}_5^3$ ? About the total number of vectors in  $\mathbb{Z}_5^3$  we see that each entry of a vector in  $\mathbb{Z}_5^3$  can independently take one of 5 possible values:  $\{0, 1, 2, 3, 4\}$ . Thus, the total number of vectors in this space is  $5^3 = 125$ . Here is a verification via Sage (to create the finite field  $\mathbb{Z}_5$  we will use the function GF)

```
# Define the finite field Z_5
F = GF(5)
# Define the 3-dimensional vector space over Z_5
V = VectorSpace(F, 3)
# Calculate the total number of vectors in the vector space
total_vectors = V.cardinality()
# Output the result
total_vectors
```

**2.C.8.** Here is a block that we can use to answer the task:

```
V=QQ^3
e1=vector([1, 0, 0])
e2=vector([0, 1, 0])
e3=vector([0, 0, 1])
Vxy=V.span([e1, e2]); print(Vxy)
Vxz=V.span([e1, e3]); print(Vxz)
Vyz=V.span([e2, e3]); print(Vyz)
```

Let us also present what Sage prints out:

```
Vector space of degree 3 and dimension 2 over Rational Field
Basis matrix:
[1 0 0]
[0 1 0]
Vector space of degree 3 and dimension 2 over Rational Field
Basis matrix:
[1 0 0]
[0 0 1]
Vector space of degree 3 and dimension 2 over Rational Field
Basis matrix:
[0 1 0]
[0 0 1]
```

**2.C.9.** For a matrix  $A$  of size  $m \times n$  the column space is defined by  $C(A) = \{Ax : x \in \mathbb{R}^n\}$ . Hence, obviously,  $C(A)$  is a subset of  $\mathbb{F}^m$ . Note that  $C(A)$  contains the zero vector, since  $A0 = 0$ . To prove that is a linear subspace of  $\mathbb{F}^m$  we should prove that  $C(A)$  is closed under vector addition and scalar multiplication. Let  $u, v \in C(A)$ . Then, there exist vectors  $x, y \in \mathbb{F}^n$  such that  $u = Ax$  and  $v = Ay$ , respectively. Thus, considering the linear combination  $\alpha u + \beta v$  of  $u, v$  for some scalars  $\alpha, \beta \in \mathbb{F}$ , we see that  $\alpha u + \beta v = \alpha(Ax) + \beta(Ay) = A(\alpha x + \beta y)$ . This shows that  $\alpha u + \beta v \in C(A)$ , and hence  $C(A)$  is a subspace of  $\mathbb{F}^m$ . Recall that the rank of  $A$  equals the number of (nonzero) pivots in elimination, hence it coincides with the maximum number of linearly independent columns (or linearly independent rows). The relation now between the dimension of  $C(A)$  and the rank of  $A$  follows. A similar analysis applies for the row space and the left null space, which we leave for practice.

**2.C.11.** The given vectors are linearly independent: The equation  $a \cdot v_1 + b \cdot v_2 = 0$  for some  $a, b \in \mathbb{R}$ , can be equivalently written in matrix form as  $(a, 0, a, 2a)^T + (0, b, 0, 0)^T = (0, 0, 0, 0)^T$ , from where we immediately get  $a = 0 = b$ . We can verify the independence of  $v_1, v_2$  in Sage as in [2.C.10](#):

```
V=QQ^4;v1=vector(QQ, [1, 0, 1, 2]); v2=vector(QQ, [0, 1, 0, 0])
V.linear_dependence([v1, v2]) ==[]
```

It follows that the linear span  $W = \text{span}_{\mathbb{Q}}\{v_1, v_2\}$  is a 2-dimensional linear subspace of  $\mathbb{Q}^4$ . As we said an alternative is given by using the command `span`. For example, adding in the previous cell the syntax

```
W=V.span([v1, v2]); W
```

verifies that  $W$  is 2-dimensional subspace of  $\mathbb{Q}^4$ . In Sage we can directly find a basis of a vector space  $V$  by the command `W.basis()`. Hence for example we can type

```
V=QQ^4
v1=vector(QQ, [1, 0, 1, 2])
v2=vector(QQ, [0, 1, 0, 0])
W=V.subspace([v1, v2])
B=W.basis();show(B)
```

By this block Sage will directly print the basis vectors  $v_1$  and  $v_2$ . Other techniques to create subspaces rely on the creation of row or column spaces of matrices, a situation that we will encounter a bit later.

**2.C.12.** (a) It is easy to see the vector equation  $av_1 + bv_2 = 0$  for some scalars  $a, b$ , implies that  $a = b = 0$ . Thus, the vectors  $v_1, v_2$  are linearly independent and automatically form a basis of  $W = \text{span}_{\mathbb{Q}}\{v_1, v_2\}$ . Therefore,  $W$  is 2-dimensional, see also the discussion in [2.3.7-2.3.10](#) for the concept of dimension of a vector space. A general element of  $W$  has the form  $w = c_1v_1 + c_2v_2 = (2c_1 + 4c_2, -c_1, 3c_1 + c_2)^T$  with  $c_1, c_2 \in \mathbb{Q}$ .

There are many ways to program Sage to verify the linear independence of  $v_1, v_2$ . Here, we present a method that uses the rank of the matrix  $A$ , which has  $v_1, v_2$  as its columns. For a built-in method in Sage to check linear independence we refer to [2.C.10](#).

```
A = matrix(QQ, [[2, 4], [-1, 0], [3, 1]])
rank_of_A = A.rank()
print(f"The rank of matrix A is: {rank_of_A}")
if rank_of_A == 2:
    print("The vectors are linearly independent.")
else:
    print("The vectors are linearly dependent.")
```

Sage's output has the form

```
The rank of matrix A is: 2
The vectors are linearly independent.
```

To introduce  $W$  in Sage we can type

```
V = VectorSpace(QQ, 3); v1 = V([2, -1, 3]); v2 = V([4, 0, 1])
W = V.subspace([v1, v2])
print("\nSubspace W spanned by v1 and v2:")
print(W)
```

An alternative to create  $W$  is to use the `span` command, as in [2.C.8](#):

```
W.span([v1, v2]); print(W)
```

For both cases Sage's output has the form

```

Subspace W spanned by v1 and v2:
Vector space of degree 3 and dimension 2 over Rational Field
Basis matrix:
[ 1  0  1/4]
[ 0  1 -5/2]

```

We will analyze the terms used in this answer in more detail soon. For now, keep only that  $W$  is a 2-dimensional subspace of  $\mathbb{Q}^3$ , over the rationals.

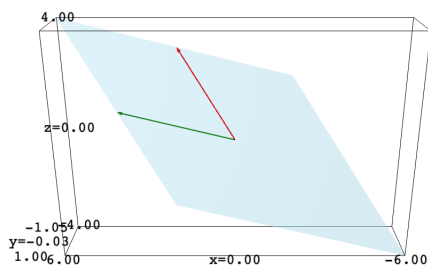
To illustrate  $W$  we can leverage Sage's 3D-graphics functions, because we are working within a three-dimensional space. We want to generate a figure displaying  $v_1$  in red,  $v_2$  in green and  $W$  as the shaded area. With this goal in mind one can execute the following block:

```

# Define the vectors
v1 = vector(QQ, [2, -1, 3])
v2 = vector(QQ, [4, 0, 1])
# Create a meshgrid for the plane using linear combinations of v1 and v2
u, v = var("u v")
plane = u*v1 + v*v2
# Define the plot range for u and v
plot_range = (-1, 1)
# Plot the plane
plane_plot = parametric_plot3d(plane,
                                (u, *plot_range),
                                (v, *plot_range),
                                color="lightblue",
                                opacity=0.5)
# Plot the vectors v1 and v2 starting from the origin
v1_plot = arrow3d((0, 0, 0), v1, color="red", thickness=0.05)
v2_plot = arrow3d((0, 0, 0), v2, color="green", thickness=0.05)
# Combine the plots
show(plane_plot + v1_plot + v2_plot, figsize=8)

```

This produces the following illustration of  $W$ :



We will meet further applications of the Sage function `parametric_plot3d` in Chapter 4.

Finally, if  $v_2$  is a multiple of  $v_1$ , then  $W$  will be reduced to a line (dimension 1).

(b) This task is a bit more challenging compared to part (a). By assumption,

$$W = \text{span}_{\mathbb{Q}}\{u_1 = (1, 1, 0, 0)^T, u_2 = (0, 1, 1, 0)^T\} = \{au_1 + bu_2 : a, b \in \mathbb{Q}\} = \{(a, a + b, b, 0)^T : a, b \in \mathbb{Q}\} \subset \mathbb{Q}^4.$$

It is easy to see that  $u_1, u_2$  are linearly independent, hence  $\dim_{\mathbb{Q}} W = 2$ , with a basis given by these vectors. For instance,

the rank of the matrix  $A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$  formed by  $u_1, u_2$  is 2. This is because by row reducing  $A$  we get

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and the final matrix has 2 pivot columns. Thus  $\text{rank}(A) = 2$  and  $u_1, u_2$  are linearly independent.

The *quotient space*  $V/W$  is the set of equivalence classes of vectors in  $V$ , under the equivalence relation  $\sim$  defined by the subspace  $W \subset V$ , i.e.,  $v_1 \sim v_2$  if and only if  $v_1 - v_2 \in W$ , where  $v_1, v_2 \in V$ . Thus, we can view  $V/W$  as  $V/W = \{[u] : u \in V\}$ , where the equivalence class of  $u \in V$  is denoted by  $[u] = \{u + w : w \in W\}$ , see 3.4.13 in Chapter 3.

The quotient space  $V/W$  inherits a vector space structure from  $V$ , with operations defined by  $[u] + [v] = [u + v]$  and  $a[u] = [au]$ , for any two vectors  $u, v$  and scalar  $a$ , try to verify this claim yourselves. It is also easy to prove that the dimension of  $V/W$  is given by  $\dim(V/W) = \dim(V) - \dim(W)$ .

For our case,  $V = \mathbb{Q}^4$  and  $\dim_{\mathbb{Q}} W = 2$ , thus  $\dim_{\mathbb{Q}} V/W = 4 - 2 = 2$ . Since  $W$  is a 2-dimensional plane in  $\mathbb{Q}^4$ , geometrically the quotient space  $V/W$  represents the remaining “directions”, after factoring out the influence of  $W$ . To introduce the quotient space  $V/W$  in Sage we can use the command `V.quotient(W)`, as follows:

```
V = VectorSpace(QQ, 4)
u1 = vector(QQ, [1, 1, 0, 0])
u2 = vector(QQ, [0, 1, 1, 0])
W = V.subspace([u1, u2])
print(V.quotient(W))
```

Let us present Sage’s output:

```
Vector space quotient V/W of dimension 2 over Rational Field where
V: Vector space of dimension 4 over Rational Field
W: Vector space of degree 4 and dimension 2 over Rational Field
Basis matrix:
[ 1  0 -1  0]
[ 0  1  1  0]
```

Notice in this answer the basis matrix refers to  $W$ . For a basis of  $V/W$  see 2.C.13.

**2.C.13.** The quotient space  $V/W$  is a 2-dimensional vector space over  $\mathbb{Q}$ , as shown in 2.C.12. Thus, a basis of  $V/W$  will consist of two vectors. These vectors can be chosen to extend the basis of  $W$  to a basis of  $V = \mathbb{Q}^4$ . For example, consider the vector  $e_3 = (0, 0, 1, 0)^T$  of the standard basis of  $\mathbb{Q}^4$ , and suppose that  $e_3 = au_1 + bu_2$  for some  $a, b \in \mathbb{Q}$ . Then we get the following system of equations  $\{a = 0, a + b = 0, b = 1\}$ , which does not admit a consistent solution for  $a, b \in \mathbb{Q}$ . This means that  $e_3 \notin W$ . Similarly we can show that  $e_4 = (0, 0, 0, 1)^T \notin W$ . Thus a basis of  $V/W$  is given by  $\underline{u} = \{e_3, e_4\}$ .

To verify with Sage that  $\underline{u}$  forms a basis of  $V/W$  we will use the `vector in subspace` command, which provides a convenient way to check if a given vector is an element of a specified subspace. Our program has the following form:

```
# Define the vectors
u1 = vector(QQ, [1, 1, 0, 0])
u2 = vector(QQ, [0, 1, 1, 0])
e3 = vector(QQ, [0, 0, 1, 0])
e4 = vector(QQ, [0, 0, 0, 1])
# Define the space W spanned by u1 and u2
V = VectorSpace(QQ, 4)
W = V.span([u1, u2])
# Check if e3 and e4 are in the span of W
def is_in_subspace(vector, subspace):
    return vector in subspace
is_e3_in_W = is_in_subspace(e3, W)
is_e4_in_W = is_in_subspace(e4, W)
# Check linear independence of e3 and e4
```

```
# Construct the matrix with e3 and e4 as rows
matrix_e3_e4 = Matrix(QQ, [e3, e4])
is_independent = matrix_e3_e4.rank() == 2
# Construct the span of e3 and e4
span_V = V.span([e3, e4])
span_check = span_V.dimension() == 2
# Display results
is_e3_in_W, is_e4_in_W, is_independent, span_check
```

Confirm yourselves that Sage's output appears as  $(\text{False}, \text{False}, \text{True}, \text{True})$ , verifying that  $e_3, e_4$  do *not* lie on  $W$  and that  $\underline{u}$  is indeed a basis of  $V/W$ . Try exploring additional examples by presenting an alternative basis.

**2.C.14.** (a) Taking  $a = b = 0$  we see that  $U$  contains the zero vector  $(0, 0, 0)^T$ . Thus, to show that  $U$  is a subspace of  $\mathbb{Z}_5^3$ , we need to prove closure under addition and scalar multiplication. Let  $u = (a_1, 2b_1, 3a_1 + 4b_1)^T$  and  $v = (a_2, 2b_2, 3a_2 + 4b_2)^T$  be arbitrary vectors of  $U$ , where  $a_i, b_i \in \mathbb{Z}_5$  for all  $i = 1, 2$ . Then, we see that

$$u + v = (a_1 + a_2, 2(b_1 + b_2), 3(a_1 + a_2) + 4(b_1 + b_2))^T \in U, \quad cu = (ca_1, 2cb_1, c(3a_1 + 4b_1)) \in U,$$

with  $c \in \mathbb{Z}_5$ . Therefore,  $U$  is a vector subspace of  $\mathbb{Z}_5^3$ .

(b) Notice that

$$\begin{pmatrix} a \\ 2b \\ 3a + 4b \end{pmatrix} = a \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix} + b \begin{pmatrix} 0 \\ 2 \\ 4 \end{pmatrix} = au_1 + bu_2, \quad \text{where } u_1 = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 0 \\ 2 \\ 4 \end{pmatrix}.$$

Thus the vectors  $u_1, u_2$  span  $U$ , i.e.,  $U = \text{span}_{\mathbb{Z}_5}\{u_1, u_2\}$ , and to obtain a basis it remains to prove linear independence.

Let  $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 3 & 4 \end{pmatrix}$  be the matrix having as columns the vectors  $u_1, u_2$ . To find the RREF of  $A$  we mention that in  $\mathbb{Z}_5$  the

inverse of 2 equals 3 and the inverse of 4 equals 4 itself. For example, for the case of 2 one needs to find an integer  $y$  such that  $2 \cdot y = 1 \pmod{5}$ . Since  $6 = 1 \pmod{5}$  our claim follows. Similarly for 4, we have  $4 \cdot 4 = 16 = 1 \pmod{5}$ . Performing elementary row operations with respect to the arithmetic in  $\mathbb{Z}_5$  (see the tables in 2.C.6), we obtain the following reduced row

echelon form:  $\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$ . For a confirmation of this matrix in Sage, add the cell

```
F = GF(5)
A=matrix(F, [[1, 0], [0, 2], [3, 4]]); A.rref()
```

It is now immediate that  $\text{rank}(A) = 2$ . Hence the vectors  $u_1, u_2$  are linearly independent. and so they form a basis of  $U$ . This implies that  $\dim_{\mathbb{Z}_5} U = 2$ . To confirm the linear independence of  $u_1, u_2$  in Sage, we use the following cell:

```
F = GF(5); V = VectorSpace(F, 3)
u1=vector(F, [1, 0, 3]); u2=vector(F, [0, 2, 4])
V.linear_dependence([u1, u2]) == []
```

By adding

```
U=span([u1, u2]); U.dimension()
```

we also get a direct verification of the dimension of  $U$ .

**2.C.16.** Considered vectors are linearly independent if and only if the vectors  $(1, -2, 0, 0)^T$ ,  $(3, 0, 1, -1)^T$ ,  $(1, -4, 1, 2)^T$ , and  $(0, 4, 8, 4)^T$  are linearly independent in  $\mathbb{R}^4$ . We see that

$$\begin{vmatrix} 1 & -2 & 0 & 0 \\ 3 & 0 & 1 & -1 \\ 1 & -4 & 1 & 2 \\ 0 & 4 & 8 & 4 \end{vmatrix} = -36 \neq 0,$$

thus the vectors are linearly independent.

**2.C.19.** In this task, first check that the given vectors  $v_1, v_2$  are linearly independent. Next, if  $A$  is the matrix having as columns these two vectors, construct the extended matrix  $(A \mid I_4)$ , where  $I_4$  is the  $4 \times 4$  identity matrix. Then compute its reduced row echelon form. From this matrix, the columns with the pivots will form the desired basis. All these in Sage can be done as follows:

```
V=QQ^4; v1=vector(QQ, [2, 0, 1, 0]); v2=vector(QQ, [1, 0, 1, 0]); L=[v1,v2]
V.linear_dependence(L)==[]
I4=identity_matrix(4); A=column_matrix(L); A1=A.augment(I4)
M=A1.rref()
[A1.column(p) for p in A1.pivots()]
```

Sage's output is here:

```
[(2, 0, 1, 0), (1, 0, 1, 0), (0, 1, 0, 0), (0, 0, 0, 1)]
```

Thus the vectors  $\{v_1, v_2, v_3 = (0, 1, 0, 0)^T, v_4 = (0, 0, 0, 1)^T\}$  form the desired basis of  $\mathbb{Q}^4$ . The linear independence of these vectors can be confirmed by adding in the previous cell the next few lines of code:

```
v3=vector(QQ, [0, 1, 0, 0]); v4=vector(QQ, [0, 0, 0, 1]); Lext=[v1,v2,v3,v4]
V.linear_dependence(Lext)==[]
```

**2.C.23.** (i) This is the projection to the  $yz$ -plane and a short computation ensures its linearity.

(ii) We see that  $F((0, 0, 0)) = (0, -3, 0) \neq (0, 0, 0)$ . However, any linear mapping maps the zero vector of its domain to the zero vector of the target space. Hence, this  $F$  is not linear.

(iii) By assumption  $a \neq 0$ . Hence, for two vector  $u = (x_1, y_1)$  and  $v = (x_2, y_2)$  lying on  $\mathbb{R}^2$ , we see that

$$F(u + v) = (x_1 + x_2 + a, y_1 + y_2) \neq F(u) + F(v) = (x_1 + a, y_1) + (x_2 + a, y_2) = (x_1 + x_2 + 2a, y_1 + y_2).$$

This shows that  $F$  is not linear.

(iv) When  $a = 0$  this is the identity map, which is obviously a linear mapping. For  $a \neq 0$  the mapping  $F(x) = x + a$  is an affine mapping. This is because the mapping  $G : \mathbb{R} \rightarrow \mathbb{R}$  defined by  $G(x) := F(x) - F(0)$  is linear, for all  $x \in \mathbb{R}$ .

(v) The mapping  $F((x, y)) = (x^2, y)$  is not linear. This is because  $F(c(x, y)) = F((cx, cy)) = (c^2x^2, cy)$  and  $cF((x, y)) = (cx^2, cy)$ . Hence in general we have  $F(c(x, y)) \neq cF((x, y))$ , except if  $c = 0$  or  $c = 1$ . Since the relation  $F(c(x, y)) = cF((x, y))$  must hold for any  $c \in \mathbb{R}$  and  $u = (x, y) \in \mathbb{R}^2$  we conclude.

(vi) The mapping  $F : \mathbb{R}^2 \rightarrow \mathbb{R}_1[t]$  with  $F((x, y)) = y + xt$  is linear. Indeed, consider two arbitrary vectors  $u = (x_1, y_1)$  and  $v = (x_2, y_2)$  on  $\mathbb{R}^2$ . Then for any two scalars  $a, b \in \mathbb{R}$  we have

$$\begin{aligned} F(au + bv) &= F((ax_1, ay_1) + (bx_2, by_2)) = F((ax_1 + bx_2, ay_1 + by_2)) = (ay_1 + by_2) + (ax_1 + bx_2)t \\ &= a(y_1 + x_1t) + b(y_2 + x_2t) = aF((x_1, y_1)) + bF((x_2, y_2)). \end{aligned}$$

(vii) Fix some  $B \in \text{Mat}_n(\mathbb{R})$ . The mapping  $F(A) = AB - BA$  is also linear, and hence an endomorphism of  $\text{Mat}_n(\mathbb{R})$ . In order to prove its linearity, consider some  $\lambda, \mu \in \mathbb{R}$  and  $A_1, A_2 \in \text{Mat}_n(\mathbb{R})$ . Then we have

$$\begin{aligned} F(\lambda A_1 + \mu A_2) &= (\lambda A_1 + \mu A_2)B - B(\lambda A_1 + \mu A_2) = \lambda A_1B + \mu A_2B - \lambda BA_1 - \mu BA_2 \\ &= \lambda(A_1B - BA_1) + \mu(A_2B - BA_2) = \lambda F(A_1) + \mu F(A_2). \end{aligned}$$

(viii) The mapping  $F(A) = A^T$  is another linear endomorphism of  $\text{Mat}_n(\mathbb{R})$ . Perform the brief computation yourselves.

(ix) Let  $u = (x_1, y_1)$  and  $v = (x_2, y_2)$  two vectors on  $\mathbb{Z}_3^2 = \mathbb{Z}_3 \times \mathbb{Z}_3$ . Then we have

$$F(u + v) = F((x_1 + x_2, y_1 + y_2)) = \begin{pmatrix} (x_1 + x_2) + (y_1 + y_2) \\ 2(x_1 + x_2) + (y_1 + y_2) \\ x_1 + x_2 \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ 2x_1 + y_1 \\ x_1 \end{pmatrix} + \begin{pmatrix} x_2 + y_2 \\ 2x_2 + y_2 \\ x_2 \end{pmatrix} = F(u) + F(v),$$

and

$$F(cu) = F((cx_1, cy_1)) = \begin{pmatrix} cx_1 + cy_1 \\ 2cx_1 + cy_1 \\ cx_1 \end{pmatrix} = c \begin{pmatrix} x_1 + y_1 \\ 2x_1 + y_1 \\ x_1 \end{pmatrix} = cF(u),$$

for all  $c \in \mathbb{Z}_3$ . Thus,  $F$  is linear.

(x) The exponential map is, of course, *not* linear because, as we will see in Chapter 5,  $e^{t_1+t_2} = e^{t_1} \cdot e^{t_2} \neq e^{t_1} + e^{t_2}$ .

**2.C.26.** Let  $\underline{e} = \{e_1 = (1, 0, 0)^T, e_2 = (0, 1, 0)^T, e_3 = (0, 0, 1)^T\}$  be the standard basis of  $\mathbb{R}^3$ . We see that

$$T(e_1) = (1, 1, 1)^T = e_1 + e_2 + e_3, \quad T(e_2) = (1, 1, 1)^T = e_1 + e_2 + e_3, \quad T(e_3) = (1, 2, 3)^T = e_1 + 2e_2 + 3e_3.$$

The columns of the matrix  $A$  corresponding to  $T$  are given by the images of the basis vectors. Thus we get

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \end{pmatrix}.$$

To work with this example in Sage, use the following code block:

```

# Define the variables and vector space
x, y, z = var("x y z")
V = W = RR^3
# Define the linear transformation
f(x, y, z) = [x + y + z, x + y + 2*z, x + y + 3*z]
T = linear_transformation(V, W, f)
# Get the matrix representation of the transformation
A = T.matrix(side="right")
# Display the matrix
show(A)

```

Verify that this confirms the matrix  $A$  obtained above.

**2.C.29.** The matrix of  $T$  with respect to the standard basis is given by  $A = \begin{pmatrix} 1 & 0 & i \\ 0 & 1 & 1 \\ i & 1 & 0 \end{pmatrix}$ . The endomorphism  $T$  will be invertible if and only if  $A$  is invertible. However,  $A$  is not of full rank, in particular,  $\text{rank}(A) = 2$  and hence  $A$  is not invertible. Alternatively, we may compute the determinant of  $A$ , which is zero,  $\det(A) = 0$ . Hence  $A$  is singular and  $T^{-1}$  does not exist. A verification is Sage is given in the usual way:

```

v1=vector([1, 0, i]); v2=vector([0, 1, 1]); v3=vector([i, 1, 0])
A=column_matrix([v1, v2, v3])
print(rank(A)); print(det(A))

```

**2.C.30.** (a) We begin with the conjugation. This maps  $1 \mapsto 1$ , and  $i \mapsto -i$ , written in the coordinates as  $(1, 0) \mapsto (1, 0)$  and  $(0, 1) \mapsto (0, -1)$ , respectively. By writing the images into the columns we obtain the matrix  $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ .

For the second map, using the basis  $\underline{u} = \{1, i\}$  we obtain  $1 \mapsto 2 + i$ ,  $i \mapsto 2i - 1$ , that is,  $(1, 0) \mapsto (2, 1)$ ,  $(0, 1) \mapsto (2, -1)$ .

Thus, the matrix of multiplication by the number  $2 + i$  under the basis  $\underline{u}$  has the form  $\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$ .

(b) We will only determine the matrix of the second linear map in the basis  $\underline{f}$ , and leave the first case for practice. Multiplication by  $(2 + i)$  gives us:  $(1 - i) \mapsto (1 - i)(2 + i) = 3 - i$ ,  $(1 + i) \mapsto (1 + 3i)$ . Coordinates  $(a, b)_{\underline{f}}$  of the vector  $3 - i$  in the basis  $\underline{f}$  are given, as we know, by the equation  $a \cdot (1 - i) + b \cdot (1 + i) = 3 + i$ , that is,  $(3 + i)_{\underline{f}} = (2, 1)$ . Analogously  $(1 + 3i)_{\underline{f}} = (-1, 2)$ . Altogether, we obtain the matrix  $\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$ . Observe that the matrices for multiplication by  $2 + i$  are the same in both bases. Why does this happen? Would the matrices be identical in both bases for multiplication by any complex number? <sup>16</sup>

**2.C.31.** To simplify the explanation, we will prove the statement for a low-dimensional case, as the idea extends easily to the general case. For example, let us consider proving that the space  $\text{Mat}_{2,3}(\mathbb{K})$  of  $2 \times 3$  matrices over  $\mathbb{K}$  is isomorphic to  $\mathbb{K}^{2 \cdot 3} = \mathbb{K}^6$ . Any element of  $\text{Mat}_{2,3}(\mathbb{K})$  has the form

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}, \quad \text{for some } a, b, c, d, e, f \in \mathbb{K}.$$

From this expression we see that the matrices

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, E_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, E_3 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, E_4 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, E_5 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, E_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

generate  $\text{Mat}_{2,3}(\mathbb{K})$ . It is also easy to see that these matrices are linearly independent, and hence they form a basis of  $\text{Mat}_{2,3}(\mathbb{K})$ . This implies that  $\dim_{\mathbb{K}} \text{Mat}_{2,3}(\mathbb{K}) = 6$ . For a verification, use Sage and its command `MatrixSpace`, as follows:

```

MS = MatrixSpace(SR, 2, 3)
MS.dimension()

```

Sage print out the number 6, which is the dimension of  $\text{Mat}_{2,3}(\mathbb{K})$  over  $\mathbb{K}$ . Adding the following command, one can also obtain the basis mentioned above:

<sup>16</sup>This similarity arises because multiplication by a complex number  $c$ , corresponds to a linear transformation that is invariant under basis changes. Essentially,  $c$  acts as a scalar multiplication on the vector space, and scalar multiplication is invariant to the choice of a basis.

```
B = MS.basis(); list(B)
```

Consider finally the mapping  $\varphi$  defined by

$$\varphi : \text{Mat}_{2,3}(\mathbb{K}) \rightarrow \mathbb{K}^6, \quad \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \mapsto (a, b, c, d, e, f)^T.$$

This is linear and maps the basis  $E_1, \dots, E_6$  to the standard basis  $e_1 = (1, 0, \dots, 0)^T, \dots, e_6 = (0, \dots, 0, 1)^T$  of  $\mathbb{K}^6$ . Hence  $\varphi$  is a linear isomorphism (you should be able to prove that a linear mapping having this property is always an isomorphism).

**2.C.32.** An isomorphism  $f : \mathbb{R}_n[x] \rightarrow \mathbb{R}^{n+1}$  is given by  $a_0 + a_1x + a_2x^2 + \dots + a_nx^n \mapsto (a_0, a_1, \dots, a_n)$ . We leave to the reader to prove that  $f$  is a well defined linear bijection. Can you specify the inverse  $f^{-1} : \mathbb{R}^{n+1} \rightarrow \mathbb{R}_n[x]$ ?

**2.C.33.** To prove the claim it is sufficient to show that  $U = \text{span}_{\mathbb{R}}\{u_1, u_2, u_3\}$  is a 3-dimensional subspace of  $\mathbb{R}^4$ , and hence that the vectors  $u_1, u_2, u_3$  are linearly independent. Indeed for some  $a, b, c \in \mathbb{R}$  we see that the matrix equation  $au_1 + bu_2 + cu_3 = 0$  is equivalent to the system  $\{a + c = 0, 2a + b = 0, b = 0\}$  having the unique solution  $a = b = c = 0$ . Or in Sage we can type

```
u1=vector([1, 2, 0, 0]); u2=vector([0, 1, 0, 1]); u3=vector([1, 0, 0, 0]); L=[u1, u2, u3]
V=RR**4; V.linear_dependence(L)==[]
```

Let  $\underline{e} = \{e_1, e_2, e_3\}$  be the standard basis of  $\mathbb{R}^3$ . Check yourselves that a linear isomorphism  $f : \mathbb{R}^3 \rightarrow U$  is defined by

$$f(e_1) = u_1, \quad f(e_2) = u_2, \quad f(e_3) = u_3.$$

**2.C.34.** A subspace of  $\mathbb{R}^3$  can be 0, 1, 2 or 3-dimensional. Thus, up to linear isomorphisms, such a subspace should be isomorphic to

- (a) the origin of  $\mathbb{R}^3$  (the single zero vector), or, (c) a plane passing through the origin of  $\mathbb{R}^3$ , or,  
 (b) a line passing through the origin of  $\mathbb{R}^3$ , or, (d)  $\mathbb{R}^3$  itself.

**2.C.36.** We can easily see that  $\underline{u}$  is a basis of  $\mathbb{C}^3$ , e.g., via Sage. To find its dual, we need to specify linear forms  $\varphi_i : \mathbb{C}^3 \rightarrow \mathbb{C}$  for  $i = 1, 2, 3$ , satisfying the relations

$\varphi_1(u_1) = 1, \varphi_1(u_2) = 0 = \varphi_1(u_3), \quad \varphi_2(u_1) = 0, \varphi_2(u_2) = 1, \varphi_2(u_3) = 0, \quad \varphi_3(u_1) = 0 = \varphi_3(u_2), \varphi_3(u_3) = 1,$   
 respectively. We may suppose that

$$\varphi_1(z_1, z_2, z_3) = \sum_{i=1}^3 a_i z_i, \quad \varphi_2(z_1, z_2, z_3) = \sum_{i=1}^3 b_i z_i, \quad \varphi_3(z_1, z_2, z_3) = \sum_{i=1}^3 c_i z_i,$$

for some complex numbers  $a_i, b_i, c_i, (i = 1, 2, 3)$ . From the first set of equations we get  $a_1 = 1, a_2 = 0$  and  $a_3 = i$ , thus  $\varphi_1(z_1, z_2, z_3) = z_1 + iz_3$ . From the second set of equations we get  $b_1 = 0, b_2 = -i$  and  $b_3 = 1$ , thus  $\varphi_2(z_1, z_2, z_3) = -iz_2 + z_3$ . Finally for  $\varphi_3$  we get  $c_1 = 0 = c_2$  and  $c_3 = -i$ , that is,  $\varphi_3(z_1, z_2, z_3) = -iz_3$ . You may try to verify alone that  $\{\varphi_1, \varphi_2, \varphi_3\}$  is indeed a basis of  $(\mathbb{C}^3)^*$ .

**2.C.37.** (a) Suppose that  $u = \sum_{i=1}^n c_i u_i$  for some real numbers  $c_i$ , for all  $i = 1, \dots, n$ . Then, by linearity, we see that

$$\varphi_1(u) = \varphi_1\left(\sum_i c_i u_i\right) = \sum_i c_i \varphi_1(u_i) = c_1 \varphi_1(u_1) + c_2 \varphi_1(u_2) + \dots + c_n \varphi_1(u_n) = c_1 \cdot 1 + c_2 \cdot 0 + \dots + c_n \cdot 0 = c_1,$$

and in an analogous way we can show that  $\varphi_i(u) = c_i$ , for all  $i = 1, \dots, n$ . Based on this remark we obtain the first relation:  $u = \sum_{i=1}^n c_i u_i = \sum_{i=1}^n \varphi_i(u) u_i$ .

(b) To prove the second relation let us apply  $\xi$  in both sides of the relation in (a). Since  $\varphi_i(u)$  are scalars for all  $i$ , i.e.,  $\varphi_i(u) \in \mathbb{K}$ , this gives

$$\xi(u) = \xi\left(\sum_i \varphi_i(u) u_i\right) = \sum_i \varphi_i(u) \xi(u_i) = \left(\sum_i \xi(u_i) \varphi_i\right)(u).$$

This relation holds for all  $u \in V$  and hence the result follows, i.e.,  $\xi = \sum_i \xi(u_i) \varphi_i$ .

**2.C.38.** If  $A = (a_{ij}), B = (b_{ij}) \in \text{Mat}_m(\mathbb{K})$  and  $c \in \mathbb{K}$ , then we have  $\text{tr}(A+B) = (a_{11} + b_{11}) + \dots + (a_{mm} + b_{mm}) = (a_{11} + \dots + a_{mm}) + (b_{11} + \dots + b_{mm}) = \text{tr}(A) + \text{tr}(B)$  and  $\text{tr}(cA) = ca_{11} + \dots + ca_{mm} = c(a_{11} + \dots + a_{mm}) = c \text{tr}(A)$ . Thus  $\text{tr}$  is a linear transformation, and since the output is a scalar, i.e.,  $\text{tr}(A) \in \mathbb{K}$ , it is a linear form of  $\text{Mat}_m(\mathbb{K})$ .

Now, for  $A, B \in \text{Mat}_m(\mathbb{K})$  we have  $AB = ((ab)_{ij})$  and  $BA = ((ba)_{ij})$ . Thus,

$$\text{tr}(AB) = \sum_{i=1}^m (ab)_{ii} = \sum_{i=1}^m \sum_{k=1}^m a_{ik} b_{ki} = \sum_{k=1}^m \sum_{i=1}^m b_{ki} a_{ik} = \sum_{k=1}^m (ba)_{kk} = \text{tr}(BA).$$



As a counterexample of the relation  $\text{tr}(AB) = \text{tr}(A)\text{tr}(B)$ , consider any non-zero matrix  $A \in \text{Mat}_m(\mathbb{R})$  and the identity matrix  $E \in \text{Mat}_m(\mathbb{R})$ . Then we see that  $\text{tr}(AE) = \text{tr}(A)$ , but  $\text{tr}(E)\text{tr}(A) = m\text{tr}(A)$ . Hence, except of the trivial case  $m = 1$  where the relation actually holds, this provides an answer.

**2.C.39.** Consider for example the endomorphism  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  represented by the matrix  $A = \begin{pmatrix} 3 & 1 \\ 0 & 2 \end{pmatrix}$  with respect to the standard basis  $\underline{e}$  of  $\mathbb{R}^2$ . Obviously,  $\text{tr}(A) = 3 + 2 = 5$ . Another basis of  $\mathbb{R}^2$  is given by  $\underline{u} = \{(1, 1)^T, (1, -1)^T\}$  (use Sage, as indicated below, to verify that these two vectors are linearly independent).

```
V=RR^2;
v1=vector(RR, [1,1])
v2=vector(RR, [1, -1])
V.linear_dependence([v1, v2]) ==[]
```

The matrix that changes from the new basis  $\underline{u}$  to the standard one  $\underline{e}$  is formed by placing the new basis vectors as columns, see also below. Let us denote this matrix by  $P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ . The matrix presentation of  $f$  in the new basis is given by  $A' = P^{-1}AP = \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}$ , see the discussion in the end of 2.3.16. We see that  $\text{tr}(A') = 5 = \text{tr}(A)$ , hence the trace is invariant under a change of basis. You may like to confirm this result in Sage (and the expression of  $A'$ ). This can be done by the usual methods, see the cell given here:

```
# Define the matrix A in the standard basis
A = Matrix(RR, [[3, 1], [0, 2]])
# Define the change of basis matrix P
P = Matrix(RR, [[1, 1], [1, -1]])
# Compute the inverse of P
P_inv = P.inverse()
# Compute the matrix representation of the endomorphism in the new basis
A_prime = P_inv * A * P
show(A_prime)
# Compute the traces
bool(A.trace() == A_prime.trace())
```

**2.C.40.** In 2.C.38 we proved the relation  $\text{tr}(AB) = \text{tr}(BA)$ . Hence, for any invertible  $m \times m$  matrix  $P$  over  $\mathbb{K}$  we obtain

$$\text{tr}(P^{-1}AP) = \text{tr}(P^{-1}(AP)) = \text{tr}((AP)P^{-1}) = \text{tr}(APP^{-1}) = \text{tr}(AE) = \text{tr}(A).$$

Thus,  $A$  and  $P^{-1}AP$  have the same trace and our claim follows. For the determinant observe that

$$\det(B) = \det(P^{-1}AP) = \det(P^{-1})\det(A)\det(P) = \det(A)\frac{\det(P)}{\det(P)} = \det(A).$$

Here, we utilized the fact that the determinant of the product of matrices equals the product of their determinants, and that for any invertible matrix  $P$ , the determinant of its inverse is given by  $\det(P^{-1}) = 1/\det(P)$ .

Now, in (a) the given matrices  $A, B$  are such that  $\text{tr}(A) = \text{tr}(B) = 4$  but  $\det(A) = 7 \neq 6 = \det(B)$ . Thus, according to our result above, the matrices  $A, B$  cannot be similar.

For the matrices  $A, B$  given in (b), we have  $\text{tr}(A) = \text{tr}(B) = 0$  and  $\det(A) = \det(B) = 0$ , hence we cannot conclude by applying our criterium. However, let  $M = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$  be an invertible  $2 \times 2$  real matrix. Thus we assume that  $\det(M) = \alpha\delta - \beta\gamma \neq 0$ , so that  $M^{-1} = \frac{1}{\alpha\delta - \beta\gamma} \begin{pmatrix} \delta & -\beta \\ -\gamma & \alpha \end{pmatrix}$ . For simplicity, we may fix  $\alpha, \beta$  so that  $\det(M) = 1$ . Then we see that

$$M^{-1}AM = \begin{pmatrix} \delta & -\beta \\ -\gamma & \alpha \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} 0 & \delta \\ 0 & -\gamma \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} \gamma\delta & \delta^2 \\ -\gamma^2 & -\gamma\delta \end{pmatrix}.$$

Therefore, in this case the matrices  $A, B$  are similar. Notice that for  $\gamma = \delta = 0$ , we have  $\det(M) = 0$ , hence we should exclude these values.

The matrices  $A, B$  given in (c) are also similar: Let us verify this in Sage via the command `B.is_similar(A)`, as follows:

```
A = matrix([[2, 1], [1, 2]])
B = matrix([[3, 0], [0, 1]])
b, P = B.is_similar(A, transformation=True)
b; show(P)
```

The output here is True, which means that  $A, B$  are indeed similar, while `show(P)` prints the matrix  $P$ , satisfying  $P^{-1}AP = B$ . A verification of this relation can be done by adding the code

```
bool(P.inverse()*A*P==B)
```

In fact, later we will explain that the diagonal matrix  $B$  in (c) consists of the eigenvalues of  $A$ , and so the similarity of  $A, B$  is related with the study of the eigenvectors of  $A$ , see also below.

**2.C.42.** To solve in Sage the task given in 2.C.41, we first introduce the vector space  $V = \mathbb{R}^3$  and confirm the linear independence of the vectors  $u_1, u_2$  and  $u_3$ . Next, we use these vectors to form the transition matrix  $T$ , which can be done easily via the `column_matrix` command. Finally, we compute the inverse of  $T$ , and display the new coordinates of  $w$  by applying the rule  $T^{-1}w$ . Here is the corresponding code:

```
V=RR^3;u1=vector([1, 1, 3])
u2=vector([1, -1, 1]);u3=vector([3, 1,5])
V.linear_dependence([u1, u2, u3])==[]
T=column_matrix([v1, v2, v3])
Tinv=T.inverse()
w=vector([1, 2, 3]); wcor=Tinv*w
show(wcor)
```

**2.C.43.** Suppose that  $u_2$  consists of the vectors  $V_1, V_2, V_3$ . By the given expression of  $T$  we have

$$V_1 = E_1 + E_2 - E_3 = (2, 0, 0)^T, \quad V_2 = E_1 + 2E_2 + E_3 = (3, 3, 2)^T, \quad V_3 = E_1 + E_2 + 2E_3 = (2, 3, 3)^T.$$

Thus  $u_2 = \{(2, 0, 0)^T, (3, 3, 2)^T, (2, 3, 3)^T\}$ . As a verification in Sage, you may type

```
V = RR**3; V1 = vector(RR, [2, 0, 0]);
V2 = vector(RR, [3, 3, 2]); V3 = vector(RR, [2, 3, 3])
V.linear_dependence([V1, V2, V3]) == [ ]
```

**2.C.44.** Again the transition matrix  $T$  for changing the basis from the basis  $\underline{f} = \{f_1, f_2, f_3\}$  to the standard basis  $\underline{e}$  can be obtained by expressing the coordinates of the vectors  $f_1, f_2, f_3$  in the standard basis as the columns of the matrix  $T$ . On the other hand, the transition matrix for changing the basis from  $\underline{e}$  to  $\underline{f}$  is the inverse of  $T$ . We compute

$$T = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} & -\frac{1}{2} \\ -\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix}.$$

Now we can derive the matrix of the mapping in the basis  $\underline{f}$ , which is given by (see also 2.2.11)

$$T^{-1}AT = \begin{pmatrix} -\frac{1}{4} & 2 & -\frac{3}{4} \\ \frac{5}{4} & 0 & \frac{7}{4} \\ \frac{3}{4} & -2 & \frac{9}{4} \end{pmatrix}.$$

A method in Sage that can be used to derive the matrix of  $F: \mathbb{R}^3 \rightarrow \mathbb{R}^3$  using built-in functions, is described in 2.E.46.

**2.C.46.** (i) As always, the expressions of the given vectors  $u_1, u_2$  are with respect to the standard basis of  $\mathbb{R}^3$ . To compute their (standard) scalar product we have

$$u_1 \cdot u_2 = \langle u_1, u_2 \rangle = u_2^T u_1 = (-1 \quad 1 \quad -\sqrt{2}) \begin{pmatrix} 1 \\ 3 \\ \sqrt{2} \end{pmatrix} = -1 + 3 - 2 = 0.$$

And the same gives the computation

$$u_2 \cdot u_1 = \langle u_2, u_1 \rangle = u_1^T u_2 = (1 \quad 3 \quad \sqrt{2}) \begin{pmatrix} -1 \\ 1 \\ -\sqrt{2} \end{pmatrix} = -1 + 3 - 2 = 0.$$

This verifies that  $\langle \cdot, \cdot \rangle$  is symmetric. Having  $u_1 \cdot u_2 = 0$  means that the vectors  $u_1, u_2$  are orthogonal each other,  $u_1 \perp u_2$ . Moreover,  $\|u_1\| = \sqrt{u_1 \cdot u_1} = \sqrt{1 + 9 + 2} = \sqrt{12} = 2\sqrt{3}$  and  $\|u_2\| = \sqrt{u_2 \cdot u_2} = \sqrt{1 + 1 + 2} = \sqrt{4} = 2$ , respectively. A verification of these computations In Sage relies on the block

```
u1 = vector([1, 3, sqrt(2)])
u2 = vector([-1, 1, -sqrt(2)])
sp1=u1.dot_product(u2);
sp2=u2.dot_product(u1); bool(sp1==sp2)
```

```
print(sp1); print(sp2)
print(norm(u1)); print(norm(u2))
```

(ii) Let us present the solution in Sage:

```
v1=vector([0, 1, 2])
v2=vector([-1, 2, 3])
spr1=v1.dot_product(v2)
print(spr1); print(norm(v1))
print(norm(v2))
```

In order to compute the angle between  $v_1, v_2$ , add the code

```
theangle = spr1/(norm(v1) * norm(v2))
arccos(theangle).n()
```

Notice the answer printed out by Sage is in radians.

**2.C.51.** It is straightforward to verify the orthonormal basis expression obtained in 2.C.48, using the function  $\text{proj}(v, u)$  developed in 2.C.48. The corresponding code block is as follows:

```
E1 = vector( [1, 1, 1, 1])
E2 = vector( [1, 1, 1, 0])
E3 = vector( [1, 1, 0, 0])
def proj (v,u):
    p = v. dot_product (u)/(norm (u)^2)*u
    return(p)
w1 = E1; w2 = E2 - proj(E2, w1)
w3 = E3 - proj(E3, w2)-proj(E3, w1)
g = [w1, w2, w3]; g
W1 = w1/norm(w1); W2 = w2/norm(w2)
W3 = w3/norm(w3); G = [W1, W2, W3]; G
```

Verify on your own that running this block, Sage will print the basis  $\{w_1, w_2, w_3\}$ , denoted by  $g$ , and the orthonormal basis  $\{\hat{w}_1, \hat{w}_2, \hat{w}_3\}$ , denoted by  $G$ .

**2.C.52.** The set of solutions of the given homogeneous linear equation is always a vector spaces. For our case, a computation shows that a basis of this vector space consists of the vectors

$$u_1 = (-1, 1, 0, 0)^T, \quad u_2 = (-1, 0, 1, 0)^T, \quad u_3 = (-1, 0, 0, 1)^T.$$

Let us now denote by  $\{v_1, v_2, v_3\}$  the orthogonal basis obtained using the Gram-Schmidt orthogonalisation process. We have  $v_1 = u_1$  and for the rest two vectors we respectively compute

$$v_2 = u_2 - \frac{u_2^T \cdot v_1}{\|v_1\|^2} v_1 = u_2 - \frac{1}{2} v_1 = \left(-\frac{1}{2}, -\frac{1}{2}, 1, 0\right)^T,$$

$$v_3 = u_3 - \frac{u_3^T \cdot v_1}{\|v_1\|^2} v_1 - \frac{u_3^T \cdot v_2}{\|v_2\|^2} v_2 = u_3 - \frac{1}{2} v_1 - \frac{1}{6} v_2 = \left(-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 1\right)^T.$$

**2.C.53.** The kernel  $\text{Ker}(\varphi)$  of  $\varphi$  is a linear subspace of  $\mathbb{R}^4$ . To specify  $\text{Ker}(\varphi)$  we need to solve the matrix equation  $Ax = 0$ , for some arbitrary vector  $x = (x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4$ . Let us use Sage to quickly perform this task:

```
A = matrix(SR, 4, 4, [[1/2, 2, -1/2, -1/2], [1, 1/2, 1, 3/2],
[2, 9/2, 0, 1/2], [2, 2, 0, 0]])
A.right_kernel()
```

Sage's output has the form

```
Vector space of degree 4 and dimension 1 over Symbolic Ring
Basis matrix: [ 1, -1, -8, 5]
```

This means that  $\text{Ker}(\varphi)$  is 1-dimensional subspace of  $\mathbb{R}^4$ , spanned by the vector  $E_1 = (1, -1, -8, 5)^T \in \mathbb{R}^4$ . For the norm of this vector we compute  $\|E_1\| = \sqrt{91}$ , hence an orthonormal basis of  $V$  consists of the vector  $\frac{1}{\sqrt{91}}E_1$ .

**2.C.54.** (i) Let  $x = (x_1, x_2, x_3, x_4)^T$  be a vector orthogonal to  $u$ , with respect to the dot product on  $\mathbb{R}^4$ . Then  $x \cdot u = x_1 + x_2 + x_3 + x_4 = 0$ . Thus, if  $U$  is the subspace of vectors orthogonal to  $u$ , then

$$U = \{(x_1, x_2, x_3, x_4)^T \in \mathbb{R}^4 : x_4 = -(x_1 + x_2 + x_3)\} = \{(x_1, x_2, x_3, -(x_1 + x_2 + x_3)^T) : x_1, x_2, x_3 \in \mathbb{R}\}.$$

It follows that  $U$  is 3-dimensional. In particular, vectors of  $U$  can be expressed as

$$(x_1, x_2, x_3, -(x_1 + x_2 + x_3))^T = x_1 E_1 + x_2 E_2 + x_3 E_3,$$

where  $E_1, E_2, E_3$  are the vectors defined by  $E_1 = (1, 0, 0, -1)^T$ ,  $E_2 = (0, 1, 0, -1)^T$ , and  $E_3 = (0, 0, 1, -1)^T$ , respectively. We easily check that these are linearly independent, and hence they provide a basis of  $U$ .

(ii) For the second task we get two conditions from orthogonality, namely:

$$x \cdot w = x_1 - x_4 = 0, \quad x \cdot z = -x_1 + x_3 = 0.$$

This means that the vector  $x = (x_1, x_2, x_3, x_4)^T$  is orthogonal both to the vectors  $w$  and  $z$  if and only if  $x_1 = x_4$  and  $x_1 = x_3$ . Therefore, if  $V$  denotes the subspace of vectors orthogonal to  $w$  and  $z$ , then we see that  $V = \{(x_1, x_2, x_1, x_1)^T : x_1, x_2 \in \mathbb{R}\}$ . It turns out that  $V$  is 2-dimensional, with a basis given by the vectors  $v_1 = (1, 0, 1, 1)^T$  and  $v_2 = (0, 1, 0, 0)^T$ .

**2.C.56.** We will prove that  $C(A)$  is orthogonal to the cokernel  $\text{Ker}(A^T)$  and leave the second case for practice. Recall that for a  $m \times n$  matrix  $A$ , both  $C(A)$  and  $\text{Ker}(A^T)$  are subspaces of  $\mathbb{R}^m$  given by  $C(A) = \{Ax : x \in \mathbb{R}^n\}$  and  $\text{Ker}(A^T) = \{y \in \mathbb{R}^m : A^T y = 0\}$ , respectively. Hence, for  $u \in C(A)$  and  $w \in \text{Ker}(A^T)$  we see that

$$\langle u, w \rangle = w^T u = w^T Ax = (A^T w)^T x = 0x = 0,$$

where the first equality is the definition of the dot product, the second occurs by replacing  $u \in C(A)$  by  $Ax$  and the third relies on the relation  $(AB)^T = B^T A^T$ . Thus  $w^T u = 0$  and  $C(A) \perp \text{Ker}(A^T)$  with respect to the dot product on  $\mathbb{R}^m$ .

Recall now from 2.C.20 that for the matrix  $A = \begin{pmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 5 \end{pmatrix}$  the column space  $C(A)$  is spanned by the vectors  $u_1 = (2, 1, 0)^T$  and  $u_2 = (4, 3, 5)^T$ , while the left null space  $\text{Ker}(A^T)$  is spanned by the vector  $u_3 = (5, -10, 2)^T$ . Hence, we see that

$$\langle u_1, u_3 \rangle = u_3^T u_1 = \begin{pmatrix} 5 & -10 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = 10 - 10 = 0, \quad \langle u_2, u_3 \rangle = u_3^T u_2 = \begin{pmatrix} 5 & -10 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 3 \\ 5 \end{pmatrix} = 20 - 30 + 10 = 0.$$

Similarly, we saw that the row space  $C(A^T)$  is the subspace  $\text{span}\{(2, 4)^T, (1, 3)^T\}$ , while  $\text{Ker}(A)$  is trivial,  $\text{Ker}(A) = \{0\}$ . Hence it is also trivial to prove that  $C(A^T) \perp \text{Ker}(A)$ .

**2.D.5.** Observe that this matrix differs from the one in 2.D.1 by just a single sign. However, in this case the characteristic polynomial of  $A$  has the form  $\lambda^3 - 6\lambda^2 + 12\lambda - 8$ , or in other words  $\chi_A(\lambda) = (\lambda - 2)^3$ . Therefore, the unique eigenvalue of  $A$  is given by  $\lambda = 2$ , with algebraic multiplicity three. The geometric multiplicity of  $\lambda$  is either one, two or three. Let us determine the eigenvectors associated to  $\lambda$ . They are the solutions of the matrix equation  $(A - 2E)x = 0$ , and a direct computation shows that the eigenspace  $V_2$  has the form  $V_2 = \text{span}_{\mathbb{R}}\{(1, -1, 0)^T, (0, 0, 1)^T\}$ . Thus, the geometric multiplicity of the unique eigenvalue  $\lambda$  of  $A$  equals to two.

**2.D.7.** Let us prove the statement for  $n = 3$  and when  $A$  is upper triangular, and similarly is treated the more general case. Thus we assume that  $A$  has the form

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix},$$

with  $a_{ij} \in \mathbb{K}$  for all  $i, j$ . Then it is easy to see that the characteristic polynomial has the form  $\chi_A(\lambda) = (a_{11} - \lambda)(a_{22} - \lambda)(a_{33} - \lambda)$ , hence the roots of  $\chi_A(\lambda)$  are the diagonal entries  $a_{11}, a_{22}, a_{33}$ . The result now follows.

**2.D.8.** (a) The first claim is based on the fact that  $\det(A)$  is the product of eigenvalues of  $A$ . On the other hand, we know that  $A$  is invertible if and only if  $\det(A) \neq 0$ . Combining these two statements we obtain that  $A$  is invertible if and only if 0 is not an eigenvalue of  $A$ .

(b) The solution arises as an application of part (a). You may like to write down the matrix  $A_n$  for small values of  $n$  and then generalize (by induction over  $n$ ). For instance, since

$$A_1 = 1, \quad A_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

we see that  $\det(A_2) = \det(A_4) = 0$  but  $\det(A_1) \neq 0$  and  $\det(A_3) \neq 0$ . Thus, an option is to compute the determinant of  $A_n$  and show that  $\det(A_n) \neq 0$  if and only if  $n$  is odd. An alternative method is suggested by the statement, and is based on the characteristic polynomial:

$$\chi_{A_n}(\lambda) = \det(A_n - \lambda E) = \begin{vmatrix} 1-\lambda & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 1 & 1-\lambda & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 1-\lambda & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & 1-\lambda & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1-\lambda & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 1-\lambda \end{vmatrix}.$$

By expanding the determinant with respect to the first column, for example, we see that  $\chi_{A_n}(\lambda) = (1-\lambda)^n + (-1)^{n-1}$ . Based on this expression we deduce that  $A_n$  is invertible if and only if  $n$  is odd (we are looking for those  $n$  such that  $\chi_{A_n}(\lambda)$  does not admit  $\lambda = 0$  as a root).

**2.D.10.** The given matrix  $A_1$  is diagonalizable, since it has two eigenvalues,  $\lambda_1 = 1$  with algebraic and geometric multiplicity 1, and  $\lambda_2 = 2$  with algebraic and geometric multiplicity 2. The eigenspace  $V_1$  is generated by the eigenvector  $(0, 1, -1/2)^T$  (or any non-zero multiple of this vector), and  $V_2$  is generated by the eigenvectors  $(1, 0, -2)^T$  and  $(0, 1, 0)^T$  (or any non-zero multiple of them). Thus, for example, the matrices  $D_1$  and  $P_1$  presented below satisfy  $A_1 = P_1 D_1 P_1^{-1}$ :

$$D_1 = \text{diag}(1, 2, 2), \quad P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 1 \\ -1 & -2 & 0 \end{pmatrix}.$$

To verify all these results via Sage, execute the following block:

```
A1=matrix(SR, [[2, 0, 0], [4, 2, 2], [-2, 0, 1]])
p(t) =A1.characteristic_polynomial(t)
show(p(t).factor())
show(p.roots())
show(A1.eigenvectors_right())
u1=vector([0, 2, -1])
u2=vector([1, 0, -2])
u3=vector([0, 1, 0])
P1=column_matrix([u1, u2, u3])
D1=diagonal_matrix([1, 2, 2])
bool(A1==P1*D1*P1.inverse())
```

For the matrix  $A_2$ , which only differs by  $A_1$  in one sign, the result is negative, in particular  $A_2$  is *not* diagonalizable. It has the same eigenvalues with  $A_1$ , i.e.,  $\lambda_1 = 1$  and  $\lambda_2 = 2$ , with algebraic multiplicity 1 and 2, respectively. However, in this case,  $\lambda_2$  has a geometric multiplicity 1 and thus is impossible to obtain a basis of eigenvectors for the matrix  $A_2$ .

**2.D.11.** Suppose that  $AA^T = A^T A = E$ . Then we see that

$$\langle \varphi_A(u), \varphi_A(w) \rangle = \langle Au, Aw \rangle = (Aw)^T Au = w^T A^T Au = w^T u = \langle u, w \rangle,$$

for any  $u, w \in V$ . Hence  $\varphi_A$  is orthogonal. For the converse, assume that  $\varphi_A$  is orthogonal. Then we have  $\langle \varphi_A(u), \varphi_A(w) \rangle = \langle Au, Aw \rangle = \langle u, w \rangle$ , or equivalently  $w^T A^T Au = w^T u$ , or  $(A^T Au)^T u = w^T u$ . Hence  $\langle u, w \rangle = \langle u, A^T Au \rangle$ . By linearity this final relation can be equivalently written as  $\langle u, (E - A^T A)w \rangle = 0$  for any  $u, w \in V$ . Recall now that  $\langle \cdot, \cdot \rangle$  is a scalar product and hence non-degenerate, which means that the condition  $\langle x, y \rangle = 0$  for any  $x$  implies  $y = 0$ . Thus we obtain  $(E - A^T A)w = 0$  for all  $w \in V$ , which yields the desired relation  $E - A^T A = 0$ , i.e.,  $A$  is orthogonal.

**2.E.3.** (i) The row operation  $R_2 \rightarrow R_2 - 3R_1$  transforms  $A$  as follows

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & -1 \\ 2 & 4 & 2 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - 3R_1} \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -7 \\ 2 & 4 & 2 \end{pmatrix}.$$

To determine the corresponding elementary matrix, which we may denote by  $E_1$ , we apply the same row operation on the  $3 \times 3$  identity matrix. This gives

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - 3R_1} \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} =: E_1.$$

To confirm this, note that the product  $E_1 A$  yields the same result as performing the row operation on  $A$ , i.e.,

$$E_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & -1 \\ 2 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -7 \\ 2 & 4 & 2 \end{pmatrix}.$$

Every elementary matrix is invertible and the inverse of an elementary matrix is also an elementary matrix. It is found by performing the reverse row operation on the identity matrix, which for our case is the row operation  $R_2 \rightarrow R_2 + 3R_1$ . Hence

we have  $E_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ , such that  $E_1 E_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = E = E_1^{-1} E_1$ .

(ii) By applying the row operation  $R_2 \rightarrow R_2 - \frac{3}{2}R_1$  we obtain  $A = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & -1 \\ 2 & 4 & 2 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - \frac{3}{2}R_1} \begin{pmatrix} 1 & 0 & 2 \\ 0 & -5 & -4 \\ 2 & 4 & 2 \end{pmatrix}$ . For

the corresponding elementary matrix we have  $E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - \frac{3}{2}R_1} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{3}{2} \\ 0 & 0 & 1 \end{pmatrix} := E_2$ . As a confirmation,

we see that the product  $E_2 A$  matches the result of the row operation on  $A$ , i.e.,  $E_2 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{3}{2} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 \\ 3 & 1 & -1 \\ 2 & 4 & 2 \end{pmatrix} =$

$\begin{pmatrix} 1 & 0 & 2 \\ 0 & -5 & -4 \\ 2 & 4 & 2 \end{pmatrix}$ . For the inverse we get  $E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{3}{2} \\ 0 & 0 & 1 \end{pmatrix}$ .

(iii) In a similar way we find that the elementary matrix representing the row operation  $R_3 \rightarrow \frac{1}{2}R_3$ , and its inverse, corresponding to the row operation  $R_3 \rightarrow 2R_3$ :

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}, \quad E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

**2.E.5.** Here is a block in Sage that confirms the expressions of  $E_1, \dots, E_8$  and of their inverses, given in 2.E.4:

```
# Define the elementary matrices
E1 = elementary_matrix(QQ, 3, row1=0, scale=1/2)
E2 = elementary_matrix(QQ, 3, row1=1, row2=0, scale=-1)
E3 = elementary_matrix(QQ, 3, row1=2, row2=0, scale=-4)
E4 = elementary_matrix(QQ, 3, row1=1, scale=-2)
E5 = elementary_matrix(QQ, 3, row1=2, scale=-1/5)
E6 = elementary_matrix(QQ, 3, row1=1, row2=2, scale=-1)
E7 = elementary_matrix(QQ, 3, row1=0, row2=2, scale=-3/2)
E8 = elementary_matrix(QQ, 3, row1=0, row2=1, scale=-1/2)
# Print the elementary matrices
show(E1, E2, E3, E4, E5, E6, E7, E8)
# Compute their inverses
E1_inv = E1.inverse()
E2_inv = E2.inverse()
E3_inv = E3.inverse()
E4_inv = E4.inverse()
E5_inv = E5.inverse()
E6_inv = E6.inverse()
E7_inv = E7.inverse()
E8_inv = E8.inverse()
# Print the inverses
show(E1_inv, E2_inv, E3_inv, E4_inv, E5_inv, E6_inv, E7_inv, E8_inv)
```

**2.E.6.** We have four unknowns  $x_1, x_2, x_3, x_4$ , and four equations. Hence the matrix  $A$  has size  $4 \times 4$ , and  $b = (3, 4, 1, 5)^T \in \mathbb{R}^4$ . Let us present the solution using Sage, and compute the reduced row echelon form of the extended matrix  $B$ :

```
B=matrix([[3, 3, 2, 1, 3], [2, 1, 1, 0, 4],
[0, 5, -4, 3, 1],[5, 3, 3, -3, 5]])
Br=B.rref(); show(Br); print(Br.pivots())
```

The output is the matrix

$$\left( \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 4 \\ 0 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 1 & 1 \end{array} \right)$$

and the quadruple  $(0, 1, 2, 3)$ . Therefore, the pivots are lying on the first four columns and the corresponding linear system admits a unique solution, given by  $x_1 = 4, x_2 = -2, x_3 = -2, x_4 = 1$ . Try to present a formal computation of the reduced row echelon form.

**2.E.9.** There are infinite many solutions represented by the vector  $(x_1 = 1 + t, x_2 = \frac{3}{2}, x_3 = t, x_4 = -\frac{1}{2})^T$ , with  $t \in \mathbb{R}$ .

**2.E.10.** This linear system has no solution. Can you explain why this claim is true?

**2.E.12.** The solution has the form  $\{(-10t, (\mu + 4)t, (3\mu - 8)t)^T : t \in \mathbb{R}\}$ .

**2.E.13.** For  $a = 0$  the system has no solution. For  $a \neq 0$  the system has infinitely many solutions.

**2.E.14.** Let us use the extended matrix of apply elementary row transformation to obtain

$$\left( \begin{array}{cccc|c} 1 & -a & -2 & b & \\ 1 & 1-a & 0 & b-3 & \\ 1 & 1-a & a & 2b-1 & \end{array} \right) \sim \left( \begin{array}{cccc|c} 1 & -a & -2 & b & \\ 0 & 1 & 2 & -3 & \\ 0 & 1 & a+2 & b-1 & \end{array} \right) \sim \left( \begin{array}{cccc|c} 1 & -a & -2 & b & \\ 0 & 1 & 2 & -3 & \\ 0 & 0 & a & b+2 & \end{array} \right).$$

Above, we first subtracted the first row from the second and the third; Then we subtracted the second row from the third. We see that the system has a unique solution (determined by backward elimination), if and only if  $a \neq 0$ . If  $a = 0$  and  $b = -2$ , we have a zero row in the extended matrix. Choosing  $x_3 \in \mathbb{R}$  as a parameter then gives infinitely many distinct solutions. For  $a = 0$  and  $b \neq -2$  the last equation  $a = b + 2$  cannot be satisfied and the system has no solution. Note that:

- For  $a = 0, b = -2$  we have infinite solutions of the form  $(x_1, x_2, x_3)^T = (-2 + 2t, -3 - 2t, t)^T$ , with  $t \in \mathbb{R}$ .
- For  $a \neq 0$  the unique solution has the form  $\left( \frac{-3a^2 - ab - 4a + 2b + 4}{a}, -\frac{2b + 3a + 4}{a}, \frac{b + 2}{a} \right)^T$ .

**2.E.17.** We compute:

$$\sigma = (1, 7)(2, 6)(5, 3), \quad \tau = (1, 6)(6, 8)(8, 7)(7, 3)(2, 4), \quad \rho = (1, 4)(4, 10)(10, 7)(7, 9)(9, 3)(2, 6)(6, 5).$$

**2.E.18.** For  $\sigma$  we compute 17 inversions, and its parity is thus odd. For  $\tau$  we enumerate 12 inversions, and hence its parity is even. Finally for  $\rho$  we enumerate 25 inversions, so its parity is odd.

**2.E.24.** A confirmation of the inverse of  $A$  and the given solution can be obtained as follows:

```
A=matrix(QQ, [[1, 1, 1, 1], [1, 1, -1, -1], [1, -1, 1, -1], [1, -1, -1, 1]])
det(A)
Ainv=A.inverse()
show(Ainv)
b=vector(QQ, [2, 3, 3, 5])
x=Ainv*b; show(x)
```

As for the computation of the algebraic complements  $A_{ij}$ , one can proceed with the following block:

```
B11=A.matrix_from_rows_and_columns ([1, 2, 3], [1, 2, 3]);show(B11)
A11=(-1)^(1+1)*det(B11); show(" A11 is equal to",A11)
B12=A.matrix_from_rows_and_columns ([1, 2, 3], [0, 2, 3]);show(B12)
A12=(-1)^(1+2)*det(B12); show(" A12 is equal to",A12)
B13=A.matrix_from_rows_and_columns ([1, 2, 3], [0, 1, 3]);show(B13)
A13=(-1)^(1+3)*det(B13); show(" A13 is equal to",A13)
B14=A.matrix_from_rows_and_columns ([1, 2, 3], [0, 1, 2]);show(B14)
A14=(-1)^(1+4)*det(B14); show(" A14 is equal to",A14)
B21=A.matrix_from_rows_and_columns ([0, 2, 3], [1, 2, 3]);show(B21)
A21=(-1)^(1+2)*det(B21); show(" A21 is equal to",A21)
B22=A.matrix_from_rows_and_columns ([0, 2, 3], [0, 2, 3]);show(B22)
A22=(-1)^(2+2)*det(B22); show(" A22 is equal to",A22)
B23=A.matrix_from_rows_and_columns ([0, 2, 3], [0, 1, 3]);show(B23)
A23=(-1)^(2+3)*det(B23); show(" A23 is equal to",A23)
B24=A.matrix_from_rows_and_columns ([0, 2, 3], [0, 1, 2]);show(B24)
A24=(-1)^(2+4)*det(B24); show(" A24 is equal to",A24)
B31=A.matrix_from_rows_and_columns ([0, 1, 3], [1, 2, 3]);show(B31)
```

```

A31=(-1)^(3+1)*det(B31); show(" A31 is equal to",A31)
B32=A.matrix_from_rows_and_columns ([0, 1, 3], [0, 2, 3]);show(B32)
A32=(-1)^(3+2)*det(B32); show(" A32 is equal to",A32)
B33=A.matrix_from_rows_and_columns ([0, 1, 3], [0, 1, 3]);show(B33)
A33=(-1)^(3+3)*det(B33); show(" A33 is equal to",A33)
B34=A.matrix_from_rows_and_columns ([0, 1, 3], [0, 1, 2]);show(B34)
A34=(-1)^(3+4)*det(B34); show(" A34 is equal to",A34)
B41=A.matrix_from_rows_and_columns ([0, 1, 2], [1, 2, 3]);show(B41)
A41=(-1)^(4+1)*det(B41); show(" A41 is equal to",A41)
B42=A.matrix_from_rows_and_columns ([0, 1, 2], [0, 2, 3]);show(B42)
A42=(-1)^(4+2)*det(B42); show(" A42 is equal to",A42)
B43=A.matrix_from_rows_and_columns ([0, 1, 2], [0, 1, 3]);show(B43)
A43=(-1)^(4+3)*det(B43); show(" A43 is equal to",A43)
B44=A.matrix_from_rows_and_columns ([0, 1, 2], [0, 1, 2]);show(B44)
A44=(-1)^(4+4)*det(B44); show(" A43 is equal to",A44)
Ain=(-1/16)*matrix(QQ, [[A11, A21, A31, A41],
[A12, A22, A32, A42], [A13, A23, A33, A43], [A14, A24, A34, A44]])
show(Ain); bool(Ain==A.inverse())
    
```

Inside this cell, the matrices  $B_{ij}$  represent the matrices  $\hat{A}_{ij}$ . The final command verifies that the matrix constructed via the algebraic complement, named here  $A_{in}$ , is the inverse of  $A$ . Or we can test our computation for the matrix  $\text{adj}(A)$ , e.g., by adding the code

```

Adtr=matrix(QQ, [[A11, A21, A31, A41], [A12, A22, A32, A42],
[A13, A23, A33, A43], [A14, A24, A34, A44]])
show(Adtr); show(A.adjugate())
bool(Adtr==A.adjugate())
    
```

**2.E.25.** To do this, we will use a slightly different method from those presented in 2.B.11. Both methods can be employed to further explore applications of Cramer's rule, such as solving linear systems with parameters. Here we will define a function to construct the matrices  $A_i$  in Cramer's notation (see 2.B.11). This will enable us to express the solution to  $Ax = b$  as  $(x_i = \det(A_i)/\det(A) : i = 1, \dots, n)$ . This function can be introduced by the block given here:

```

def column_replace (M, column ,u):
    n1=M.nrows ()
    n2=M.ncols ()
    P = matrix (n1 ,n2)
    for i in range (n1):
        for j in range (n2):
            P[i,j]=M[i,j]
    for i in range (n1):
        P[i,column -1]= u[i ,0]
    return P
    
```

To solve our task, we can now add the following cell:

```

A= matrix ([[1 ,1 ,1 ,1] , [1 , 1 , -1, -1] , [1 ,-1, 1, -1] , [1 ,-1 ,-1, 1]])
b= matrix (4 ,1 , [2 ,3 ,3 ,5])
A1= column_replace (A ,1,b)
A2= column_replace (A ,2,b)
A3= column_replace (A ,3,b)
A4= column_replace (A ,4,b)
x= matrix (4 ,1 , [det(A1)/det(A), det(A2)/det(A), det(A3)/det(A), det(A4)/det(A)])
show(x)
    
```

Executing this block Sage returns the solution obtained above, i.e.,  $(\frac{13}{4}, -\frac{3}{4}, -\frac{3}{3}, \frac{1}{4})^T$ .

**2.E.27.** One may compute first  $F^{-1}$ . Then, based on the formula  $F^{-1} = \frac{1}{\det(F)}\text{adj}(F)$  we deduce that

$$\text{adj}(F) = (\alpha\delta - \beta\gamma) F^{-1} = \begin{pmatrix} \delta & -\beta & 0 \\ -\gamma & \alpha & 0 \\ 0 & 0 & \alpha\delta - \beta\gamma \end{pmatrix},$$

for all  $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ .



**2.E.28.** The corresponding adjoint matrices have the form

$$\operatorname{adj}(A) = \begin{pmatrix} 1 & 1 & -2 & -4 \\ 0 & 1 & 0 & -1 \\ -1 & -1 & 3 & 6 \\ 2 & 1 & -6 & -10 \end{pmatrix}, \quad \operatorname{adj}(B) = \begin{pmatrix} 6 & -2i \\ -3 + 2i & 1 + i \end{pmatrix}.$$

**2.E.29.** We can easily verify that  $A$  satisfies  $A^2 = A$ . In Sage give the cell

```
A=matrix([[2, -3, -5], [-1, 4, 5], [1, -3, -4]]); bool(A==A*A)
```

Sage returns True. As for the product  $AB$ , we get  $AB = A(A - I) = A^2 - A = A - A = 0$ . Moreover, for  $B^2$  we see that

$$B^2 = (A - I)(A - I) = A^2 - 2A - I = I - A = -B = \begin{pmatrix} -1 & 3 & 5 \\ 1 & -3 & -5 \\ -1 & 3 & 5 \end{pmatrix}.$$

In Sage a verification of these results it is very fast. For example add the following syntax in the previous cell:

```
E=matrix([[1, 0, 0], [0, 1, 0], [0, 0, 1]]); B=A-E; show(A*B); show(B^2)
```

Finally, let us use Sage again to compute  $\det(A)$ , where we just need to add the command  $\det(A)$ . It gives  $\det(A) = 0$ , i.e.,  $A$  is a *singular* matrix. Try to present a formal computation of  $\det(A)$ . In fact, except of the identity matrix, all the other idempotent matrices have determinant zero (see 2.E.61 below).

**2.E.30.** Both relations are true. To prove this, we will use Cauchy's theorem, which states that  $\det(AB) = \det(A)\det(B)$  (see 2.2.7). Since  $\det(PP^{-1}) = \det(E) = 1$  we obtain

$$\det(B) = \det(PAP^{-1}) = \det(P)\det(A)\det(P^{-1}) = \det(P)\det(P^{-1})\det(A) = \det(PP^{-1})\det(A) = \det(A).$$

Hence,  $\det(B) = \det(A)$  and we also obtain  $\det(A^{-1}B) = \det(A^{-1})\det(B) = \det(A^{-1})\det(A) = \det(A^{-1}A) = 1$ .

**2.E.32.** This task is solved in a similar way with 2.E.31, and you can verify that the desired coordinates are given by  $(2 + \frac{1}{\sqrt{3}}, 2 - \frac{1}{\sqrt{3}})$ .

**2.E.33.** Let us write

$$(5, 1, 11)^T = p(3, 2, 2)^T + q(2, 3, 1)^T + r(1, 1, 3)^T$$

for some  $p, q, r \in \mathbb{R}$ . Solving the corresponding linear system we obtain a unique solution given by  $p = 2$ ,  $q = -2$  and  $r = 3$ .

**2.E.34.** We see that the vectors  $u_1, u_2, u_3$  are linearly dependent, whenever at least one of the following conditions is satisfied:  $a = b = 1$ , or  $a = c = 1$ , or  $b = c = 1$ .

**2.E.35.** It turns out that the given vectors are linearly independent. In order to prove our assertion, you may try to apply the method presented in 2.C.16.

**2.E.36.** It is easy to check that adding the polynomial  $x$  gives a basis of  $\mathbb{R}_3[x]$ .

**2.E.37.** For both cases we have to check the linear independence of the given vectors  $u_1, u_2, u_3$ .

(a) In the first case we compute  $\dim U = 2$ , for  $t \in \{1, 2\}$ , otherwise we have  $\dim U = 3$ .

(b) In the second case we compute  $\dim U = 2$  for  $t \neq 0$ , and  $\dim U = 1$  for  $t = 0$ .

**2.E.39.** According to the definition of intersection, the vectors in the intersection are in both, the span of the vectors  $(1, 1, -3)$ ,  $(1, 2, 2)$ , as well as in the span of the vectors  $(1, 1, -1)$ ,  $(1, 2, 1)$ ,  $(1, 3, 3)$ . We see that  $U$  is spanned by two linearly independent vectors, and hence  $U$  is a plane in  $\mathbb{R}^3$ . Next,  $V$  is spanned by three vectors, but these are linearly dependent:

$$\begin{vmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ -1 & 1 & 3 \end{vmatrix} = \begin{vmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \end{vmatrix} = 0.$$

Therefore,  $V$  is also a plane. If the vector  $(x_1, x_2, x_3)$  lies in  $U$ , then  $(x_1, x_2, x_3) = \lambda(1, 1, -3) + \mu(1, 2, 2)$ , for some scalars  $\lambda, \mu$ . Similarly, if the vector  $(x_1, x_2, x_3)$  lies in  $V$ , then  $(x_1, x_2, x_3) = \alpha(1, 1, -1) + \beta(1, 2, 1) + \gamma(1, 3, 3)$ , for some scalars  $\alpha, \beta, \gamma$ . This gives a system of six equations with eight unknowns. Solving such a system can be quite cumbersome, but we can make some simplifications. First notice that the subspace  $U$  is described by the following three equations:

$$x_1 = \lambda + \mu, \quad x_2 = \lambda + 2\mu, \quad x_3 = -3\lambda + 2\mu.$$

Solving this system of equations with respect to  $\lambda$  and  $\mu$ , or alternatively by eliminating  $\lambda$  and  $\mu$  from these equations, one obtains the single equation  $8x_1 - 5x_2 + x_3 = 0$ , which we may use to replace the first three. Now notice that the subspace  $V$  is describe by the following three equations:

$$x_1 = \alpha + \beta + \gamma, \quad x_2 = \alpha + 2\beta + \gamma, \quad x_3 = -\alpha + \beta + 3\gamma.$$

Solving this systems of equations with respect to  $\alpha$ ,  $\beta$  and  $\gamma$ , or alternatively by eliminating  $\alpha$ ,  $\beta$  and  $\gamma$  from these equations, we obtain the single equation  $3x_1 - 2x_2 + x_3 = 0$ , which we can use to describe  $V$ . Hence, it is now straightforward that after introducing a new parameter  $t$ , we can express intersection by  $(x_1, x_2, x_3) = t \cdot (3, 5, 1)$ .

**2.E.40.** The answer is given by the matrix  $\begin{pmatrix} 5/6 & -1/6 & 1/3 \\ -1/6 & 5/6 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$ .

**2.E.41.** The answer is given by the matrix  $\begin{pmatrix} 5/9 & 2/9 & -4/9 \\ 2/9 & 8/9 & 2/9 \\ -4/9 & 2/9 & 5/9 \end{pmatrix}$ .

**2.E.42.** (a) The vector that determines the subspace  $U$  is perpendicular to each of the three vectors that generate  $W$ . Thus the subspaces are orthogonal and the first claim holds.

(b) It is *not* true that  $\mathbb{R}^4 = U \oplus W$ . This is because the subspace  $W$  is only two-dimensional, since  $(-1, 0, -1, 2)^T = (-1, 0, 1, 0)^T - 2(0, 0, 1, -1)^T$ .

**2.E.43.** Let

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

be the  $4 \times 3$  matrix induced by the vectors  $E_1, E_2, E_3$ . Recall that the maximum number of linearly independent row vectors in  $A$  coincides with the rank of  $A$ . Because the determinant of the first three rows equals to  $1 \neq 0$ , the rank of  $A$  equals to 3,  $\text{rank}(A) = 3$ . Hence the vectors  $E_1, E_2, E_3$  are linearly independent. Recall that to verify this conclusion by Sage it suffices to add the block

```
V = RR^4
E1 = vector(RR, [1, 0, 0, 1])
E2 = vector(RR, [0, 1, 0, 1])
E3 = vector(RR, [1, 2, 1, 0])
V.linear_dependence([E1, E2, E3]) == []
```

On the other hand, for the rank of  $A$  the appropriate cell is

```
A = matrix(SR, [[1, 0, 1], [0, 1, 2], [0, 0, 1], [1, 1, 0]])
A.rank()
```

In the first case Sage's output is True, while in the second Sage prints out the desired number 3.

Since  $E_1, E_2, E_3$  are linearly independent and span  $W$ , they form a basis of  $W$ . We will use this basis to obtain an orthogonal basis by applying the usual Gram–Schmidt procedure. Set

$$w_1 = E_1, \quad w_2 = E_2 - \frac{\langle E_2, w_1 \rangle}{\|w_1\|^2} w_1, \quad w_3 = E_3 - \frac{\langle E_3, w_1 \rangle}{\|w_1\|^2} w_1 - \frac{\langle E_3, w_2 \rangle}{\|w_2\|^2} w_2.$$

Since  $\langle E_2, w_1 \rangle = 1$  and  $\|w_1\|^2 = 2$  this gives  $w_2 = (-\frac{1}{2}, 1, 0, \frac{1}{2})^T$ . We also compute  $\langle E_3, w_1 \rangle = 1$ ,  $\langle E_3, w_2 \rangle = \frac{3}{2}$  and  $\|w_2\|^2 = \frac{3}{2}$ . Hence  $w_3 = E_3 - \frac{1}{2}E_1 - w_2$ , that is  $w_3 = (\frac{1}{2}, 1, 1, -1)^T$ , with  $\|w_3\|^2 = \frac{13}{4}$ . Thus an orthonormal basis of  $W$  consists of the vectors

$$\hat{w}_1 = \frac{w_1}{\|w_1\|} = \left(\frac{1}{\sqrt{2}}, 0, 0, \frac{1}{\sqrt{2}}\right)^T, \quad \hat{w}_2 = \frac{w_2}{\|w_2\|} = \left(\frac{-1}{\sqrt{6}}, \sqrt{\frac{3}{2}}, 0, \frac{1}{\sqrt{6}}\right)^T, \quad \hat{w}_3 = \frac{w_3}{\|w_3\|} = \left(\frac{1}{\sqrt{13}}, \frac{2}{\sqrt{13}}, \frac{2}{\sqrt{13}}, \frac{-2}{\sqrt{13}}\right)^T.$$

By definition, we have  $\mathbb{R}^4 = W \oplus W^\perp$  and hence  $\dim W^\perp = \dim \mathbb{R}^4 - \dim W = 1$ . Choosing any vector of the orthonormal basis of  $\mathbb{R}^4$ , say  $e_2 = (0, 1, 0, 0)$ , we extend the above basis to a basis to  $\mathbb{R}^4$ , namely  $\{Z_1 = \hat{w}_1, Z_2 = \hat{w}_2, Z_3 = \hat{w}_3, Z_4 = e_2\}$  (we encourage the reader to verify this statement in Sage). By applying the Gram–Schmidt method to this basis we obtain an orthogonal basis  $\{\hat{Z}_1, \dots, \hat{Z}_4\}$  of  $\mathbb{R}^4$ , given by  $\hat{Z}_j = Z_j = \hat{w}_j$  for  $j = 1, 2, 3$  and

$$\hat{Z}_4 = Z_4 - \frac{\langle Z_4, \hat{w}_1 \rangle}{\|\hat{w}_1\|^2} \hat{w}_1 - \frac{\langle Z_4, \hat{w}_2 \rangle}{\|\hat{w}_2\|^2} \hat{w}_2 - \frac{\langle Z_4, \hat{w}_3 \rangle}{\|\hat{w}_3\|^2} \hat{w}_3.$$

It follows that  $W^\perp = \text{span}_{\mathbb{R}}\{\hat{Z}_4\}$ , and the explicit computation of  $\hat{Z}_4$  is left to the reader.

**2.E.47.** Using the Gram-Schmidt orthogonalization process we obtain the basis:  $\{(1, 1, 1, 1)^T, (1, 1, 1, -3)^T, (-2, 1, 1, 0)^T\}$ .

**2.E.48.** For example, one can obtain the orthogonal bases  $\{(1, 0, 1, 0)^T, (0, 1, 0, -7)^T\}$  for the first part, and  $\{(1, 2, 2, -1)^T, (2, 3, -3, 2)^T, (2, -1, -1, -2)^T\}$  for the second part.

**2.E.49.** The solution is  $a = 9/2, b = -5$  (since  $1 + b + 4 + 0 + 0 = 0$  and  $1 - b + 0 + 3 - 2a = 0$ ).

**2.E.50.** The orthogonal complement  $V^\perp$  is the set of all scalar multiples of the vector  $(4, 2, 7, 0)^T$ .

**2.E.51.** Obviously, there are infinitely many possible extensions. For example, a very simple one is given by the set

$$\{u_1, u_2, u_3 = (1, 0, 0, -1)^T, u_4 = (1, 0, -1, 1)^T\}.$$

**2.E.52.** The basis consists only from the vector  $u_1 = (3, -7, 1, -5, 9)^T$  (or any non-zero scalar multiple of  $u_1$ ).

**2.E.53.** The answers are given as follows:

(a)  $W^\perp = \text{span}_{\mathbb{R}} \{(1, 0, -1, 1, 0)^T, (1, 3, 2, 1, -3)^T\}$ .

(b)  $W^\perp = \text{span}_{\mathbb{R}} \{(1, 0, -1, 0, 0)^T, (1, -1, 1, -1, 1)^T\}$ .

**2.E.54.** Obviously, any squared matrix  $A$  can be written as  $A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$ . Setting  $A_s = \frac{1}{2}(A + A^T)$  and  $A_a = \frac{1}{2}(A - A^T)$ , respectively, it is clear that

$$A_s^T = \frac{1}{2}(A + A^T)^T = A_s, \quad A_a^T = \frac{1}{2}(A - A^T)^T = -A_a.$$

This proves the first statement, i.e.,  $A = A_s + A_{\text{skew}}$ . For the given matrix  $A$  we compute

$$A_s = \frac{1}{2} \left\{ \begin{pmatrix} 1 & 0 & 2 \\ 6 & 3 & 0 \\ 2 & 2 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 6 & 2 \\ 0 & 3 & 2 \\ 2 & 0 & 4 \end{pmatrix} \right\} = \begin{pmatrix} 1 & 3 & 2 \\ 3 & 3 & 1 \\ 2 & 1 & 4 \end{pmatrix}$$

and

$$A_a = \frac{1}{2} \left\{ \begin{pmatrix} 1 & 0 & 2 \\ 6 & 3 & 0 \\ 2 & 2 & 4 \end{pmatrix} - \begin{pmatrix} 1 & 6 & 2 \\ 0 & 3 & 2 \\ 2 & 0 & 4 \end{pmatrix} \right\} = \begin{pmatrix} 0 & -3 & 0 \\ 3 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Observe that a skew-symmetric matrix, as  $A_a$ , has zeros in the diagonal. In Sage, to find  $A_s$  and  $A_a$  we can proceed with the cell

```
A=matrix([[1, 0, 2], [6, 3, 0], [2, 2, 4]]); show(A)
As=(A+A.transpose())/2; Aa=(A-A.transpose())/2; show(As); show(Aa)
As.is_symmetric() # this command is used to verify that A_s is symmetric
```

Notice an alternative to compute the transpose of a matrix  $A$  in Sage, is the shortcut  $A.T$ , a notation which is common in some other programming environments like NumPy. Hence for example to determine the matrix  $A_s$  in Sage one could type

```
A=matrix([[1, 0, 2], [6, 3, 0], [2, 2, 4]])
As=(A+A.T)/2; print(As)
```

**2.E.55.** (a) The assertions are obvious:

$$(A + A^T)^T = A^T + (A^T)^T = A^T + A = A + A^T, \quad (A - A^T)^T = A^T - (A^T)^T = A^T - A = -(A - A^T).$$

(b) By assumption  $A$  is skew-symmetric, i.e.,  $A = -A^T$  and  $n$  is odd. Recall that  $\det(X) = \det(X^T)$  and  $\det(cX) = c^n \det(X)$  for any  $n \times n$  matrix  $X$ , see 2.B.8. Thus,  $\det(A) = \det(A^T) = \det(-A) = (-1)^n \det(A) = -\det(A)$  which implies that  $2 \det(A) = 0$ , i.e.,  $\det(A) = 0$ .

(c) The proof for the claim about the trace is also easy: Since a skew-symmetric matrix  $A = (a_{ij})$  satisfies  $a_{ij} = -a_{ji}$  for any  $i, j$ , we have  $a_{ii} = 0$  for all  $i = j$ . Thus, all the diagonal entries of a skew-symmetric matrix are zero.

**2.E.58.** The given identity occurs easily by (\*), which for  $\lambda = 1$  gives  $\chi_A(1) = \det(A - E) = -1 + \text{tr}(A) + c_1 + \det(A)$ , so we can replace  $c_1$  in (\*) by  $\det(A - E) + 1 - \text{tr}(A) - \det(A)$ .

For the given matrix  $A$  we compute  $\text{tr}(A) = 12$  and  $\det(A) = 60$ . Hence an application gives

$$\chi_A(\lambda) = \det(A - \lambda E) = -\lambda^3 + 12\lambda^2 - 47\lambda + 60.$$

Using Sage we find that  $\chi_A(\lambda)$  has three roots:  $\lambda_1 = 3, \lambda_2 = 4$ , and  $\lambda_3 = 5$ . These are the eigenvalues of  $A$ , all with algebraic multiplicity one. Try to solve the equation  $\chi_A(\lambda) = 0$  in a formal way. For a description in Sage of the given expression of the characteristic polynomial, but also for a verification of the determinant and trace, type

```
A=matrix(SR, [[32, -67, 47], [7, -14, 13], [-7, 15, -6]])  
print(A.trace( )); print(A.det( ))
```

```
p(t) = A.characteristic\_polynomial(t)
show(p(t)); show(p.roots())
```

**2.E.59.** Possible examples illustrated by the matrices  $A, B$  and  $C$  provided below, which correspond to the cases (i), (ii) and (iii), respectively:

$$A = \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 7 \end{pmatrix}, \quad B = \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 1 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 7 \end{pmatrix}, \quad C = \begin{pmatrix} 6 & 0 & 0 & 0 \\ 0 & 7 & 1 & 0 \\ 0 & 0 & 7 & 1 \\ 0 & 0 & 0 & 7 \end{pmatrix}.$$

Explain the reasoning behind the derivation of these matrices and then explore additional solutions.

**2.E.60.** There is a triple eigenvalue  $-1$ . The corresponding eigenspace is spanned by the eigenvectors  $(1, 0, 0)^T$  and  $(0, 2, 1)^T$ , hence it is 2-dimensional (over  $\mathbb{R}$ ).

**2.E.65.** The matrix has a double eigenvalue  $-1$ , and its associated eigenspace is spanned over  $\mathbb{R}$  by the vectors  $(2, 0, 1)^T$  and  $(1, 1, 0)^T$ . Further, the matrix has  $0$  as the eigenvalue, with eigenvector  $(1, 4, -3)^T$ . Combining this with the statements in 2.E.63, we deduce that the mapping induced by the given matrix  $M$  is an axial symmetry through the line induced by the last vector, composed with the projection on the plane perpendicular to the last vector. Verify that this plane is given by the equation  $x + 4y - 3z = 0$ .

**2.E.66.** (i) Let us present the computations in Sage, where we can use the command `A.commutator(B)` to compute the commutator  $[A, B] = AB - BA$  of two squared matrices  $A, B$ . The corresponding block is here:

```
s1=matrix(SR, [[0, 1], [1, 0]])
s2=matrix(SR, [[0, -i], [i, 0]])
s3=matrix(SR, [[1, 0], [0, -1]])
show(s1.commutator(s2)); bool(s1.commutator(s2)==2*i*s3)
show(s1.commutator(s3)); bool(s1.commutator(s3)==-2*i*s2)
show(s2.commutator(s3)); bool(s2.commutator(s3)==2*i*s1)
```

Execute this block to read Sage's output.

(ii) To verify the claims posed in the second part one can proceed by adding in the previous cell the code

```
bool(s1*s1==identity_matrix(2))
bool(s2*s2==identity_matrix(2))
bool(s3*s3==identity_matrix(2))
ev1=s1.eigenvalues()
ev2=s2.eigenvalues()
ev3=s3.eigenvalues()
bool(ev1==ev2); bool(ev1==ev3); bool(ev2==ev3); show(ev1)
```

Let us also verify the claim for the eigenvalues by a formal computation, e.g., for  $\sigma_1$ . The point is to compute its characteristic polynomial  $\chi_{\sigma_1}(\lambda) = \det(\sigma_1 - \lambda E)$ , where here  $E = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ . We have

$$\sigma_1 - \lambda E = \begin{pmatrix} -\lambda & 1 \\ 1 & -\lambda \end{pmatrix} \implies \chi_{\sigma_1}(\lambda) = \begin{vmatrix} -\lambda & 1 \\ 1 & -\lambda \end{vmatrix} = \lambda^2 - 1.$$

Thus there are two roots,  $\lambda_{\pm} = \pm 1$ . Similarly for  $\sigma_2$  and  $\sigma_3$ . Try to you determine the corresponding eigenvectors.

(iii) The claim in this part can be proved similarly, and left as an exercise.

**2.E.67.** This is direct and left for practice.

## Linear models and matrix calculus

where are the matrices useful?  
– basically almost everywhere...



### A. Linear optimization

The idea of maximizing or minimizing a linear function subject to linear constraints arises naturally in many fields. For instance we may want to maximize profits, minimize costs, etc. Linear programming builds upon notions from linear algebra discussed in Chapter 2 and is a dynamic tool for solving linear optimization problems subject to linear constraints.<sup>1</sup>

In a linear programming problem (LP problem in short) we seek for a vector on some Euclidean space maximizing (or minimizing) the value of some linear functional, among all vectors satisfying a given system of linear constraints that govern the process. In 3.1.2 this linear functional is called the objective function.

**3.A.1.** To illustrate the idea let us begin with LP problems in two dimensions, where a solution occurs graphically. As usual, we view vectors  $x$  on  $\mathbb{R}^2$  as column matrices  $(x_1, x_2)^T$  with  $x_1, x_2 \in \mathbb{R}$ . We want to maximize the value of the linear function  $h(x_1, x_2) = x_1 + x_2$  (objective function) subject to the constraints

$$x_1 + 2x_2 \leq 3, \quad 2x_1 + x_2 \leq 3, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

<sup>1</sup>The name "programming" refers to process planning and was introduced before the computers were invented, cf. the footnote on page 195.

We have already developed a useful package of tools and it is time to show some applications of matrix calculus. The first three parts of this chapter are independent and the readers more interested in the theory might skip any of them and continue with the fourth part straight ahead.

It might seem that the assumption of linearity of relations between quantities is too restrictive. But this is often not so. In real problems, linear relations may appear directly. A problem may be solved as a result of an iteration of many linear steps. If this is not the case, we may still use this approach at least to approximate real non-linear processes.

We should also like to compute with matrices (and linear mappings) as easily as we can compute with scalars. In order to do that, we prepare the necessary tools in the second part of this chapter. We also present a useful application of matrix decompositions to the pseudoinverse matrices, which are needed for numerical mastery of matrix calculus.

We try to illustrate all the phenomena with rather easy problems. Still some parts of this chapter are perhaps difficult for first reading. This in particular concerns the very first part providing some glimpses towards the linear optimization (linear programming), and the third part devoted to iterated processes (the Frobenius-Perron theory).

The rest of the chapter comes back to some more advanced parts of the matrix calculus (the Jordan canonical form, decompositions, and pseudo-inverses of matrices). The reader should feel free to move forward if getting lost.

### 1. Linear optimization

The simplest linear processes are given by linear mappings  $\varphi : V \rightarrow W$  on vector spaces. As we can surely imagine, the vector  $v \in V$  can represent the state of some system we are observing, while  $\varphi(v)$  gives the result after some process is realized.

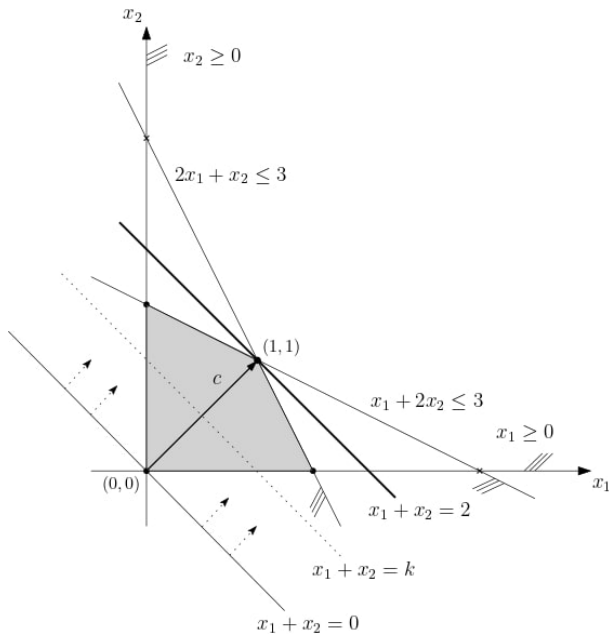
If we want to reach a given result  $b \in W$  of such a process, we solve the problem

$$\varphi(x) = b$$

for some unknown vector  $x$  and a known vector  $b$ .

In fixed coordinates we then have the matrix  $A$  of a mapping  $\varphi$  and coordinate expression of the vector  $b$ . We have mastered such problems in the previous chapter. Now we

**Solution.** The inequalities are graphically represented by half-planes in  $\mathbb{R}^2$ , and the intersection of these half-planes determines the “feasible region” of solutions. Thus, a feasible solution is any point in  $\mathbb{R}^2$  satisfying all the given constraints. The non-negativity constraints  $x_i \geq 0$  for  $i = 1, 2$  clearly restrict all the feasible solutions to the first quadrant. Therefore, the feasible region is the shaded (hatched) area in the diagram below. While it is possible to construct linear programming (LP) problems with an empty feasible region, we are primarily interested in those with a non-empty feasible region.



In a linear maximization problem, an “optimal solution” is a point within the feasible region that maximizes the objective function’s value.

For our example, an optimal solution appears at a unique vertex. To see this, let us consider the vector  $c = (1, 1)^T$ , such that  $c^T x = x_1 + x_2$ . Consider also the objective function line  $x_1 + x_2 = k$  ( $k \in \mathbb{R}$ ), perpendicular to the vector  $c$ . As illustrated in the figure above, we now begin moving this line upwards in the direction of the vector  $c$ , maintaining the same slope, without leaving the feasible region. By doing so, we can determine the point where the moving line intersects the feasible region for the last time, that is, at the point  $(1, 1)$ . This point provides the maximum value  $k = 2$  for the objective function  $h$ . □

*On the other hand, there exist LP problems with infinitely many optimal solutions. Nevertheless, at least one of these optimal solutions must occur at a vertex of the feasible region. Let us describe such an example.*

**3.A.2.** Consider the problem of maximizing the functional  $h(x_1, x_2) = 2.5x_1 + x_2$  under the constraints

$$3x_1 + 5x_2 \leq 15, \quad 5x_1 + 2x_2 \leq 10, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

draw more interesting conclusions in the setup of *linear optimization models* (called also *linear programming*).

**3.1.1. Linear optimization.** In the practical column, the previous chapter started with a painting problem, and we shall continue here in a similar way. Imagine that our very specialized painter in a black&white world is willing to paint facades of either small family houses or of large public buildings, and that he (of course) uses only black and white colours. He can arbitrarily choose proportions between  $x_1$  units of area for the small houses or  $x_2$  units for the large buildings. Assume that his maximal workload in a given interval of time is  $L$  units of area, his net income (that is, after subtracting the costs) is  $c_1$  per unit of area for small houses and  $c_2$  per unit of area for large buildings. Furthermore, he has only  $W$  kg of white colour and  $B$  kg of black colour at his disposal. Finally, a unit of area for small houses requires  $w_1$  kg of white colour and  $b_1$  kg of black colour. For large buildings the corresponding values are  $w_2$  and  $b_2$ .

If we write all this information as inequalities, we obtain the conditions

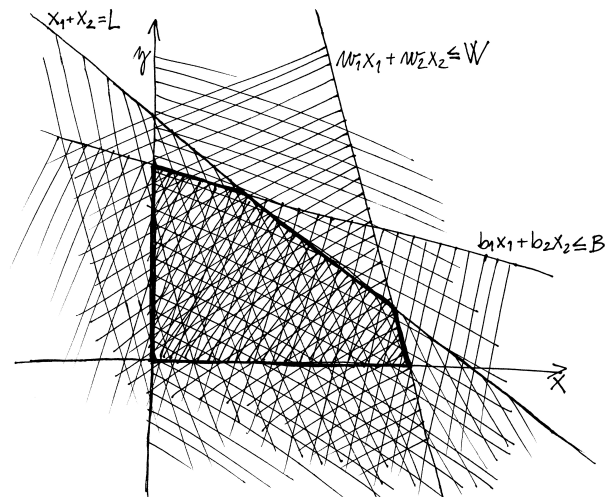
- (1)  $x_1 + x_2 \leq L$
- (2)  $w_1x_1 + w_2x_2 \leq W$
- (3)  $b_1x_1 + b_2x_2 \leq B$ .

The total net income of the painter, which is the following linear form  $h$ ,

$$h(x_1, x_2) = c_1x_1 + c_2x_2,$$

is to be maximized.

Each of the given inequalities clearly determines a half-plane in the plane of the variables  $(x_1, x_2)$ , bounded by a line given by the corresponding equality, and we must also assume that both  $x_1$  and  $x_2$  are non-negative real numbers (because the painter cannot paint negative areas). Thus we have constraints for the values  $(x_1, x_2)$  – either the constraints are unsatisfiable, or they allow points inside a polygon with at most five vertices. See the diagram.



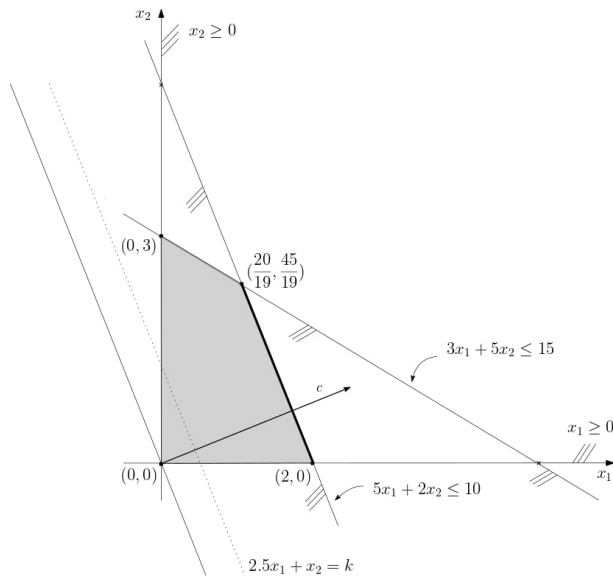
the axis in the diagram should be called  $x_1$  and  $x_2$ ! Add the line of constant value for  $h$ , best through one of the vertices with hand-written description “optimal constant value of  $h$ ”

**Solution.** Obviously, we have  $h = c^T x$  where  $c = (2.5, 1)^T$  and the feasible region is the shaded polygon in the figure below. All the points on the thick side of the polygon are optimal solutions. This is because  $h$  is parallel to the line  $5x_1 + 2x_2 = 10$ , induced by the second constraint. Indeed, we see that both the points  $(2, 0)$  and  $(\frac{20}{19}, \frac{45}{19})$  are optimal solutions to the problem, with  $h(0, 2) = h(\frac{20}{19}, \frac{45}{19}) = 5$ . The line segment  $p$  joining these points has the form

$$(x_1, x_2) = t(2, 0) + (1-t)(\frac{20}{19}, \frac{45}{19}) = (\frac{20}{19} + \frac{18}{19}t, \frac{45}{19} - \frac{45}{19}t)$$

with  $t$  ranging from 0 to 1. Then, any point  $(x_1, x_2)$  on  $p$  maximizes  $h$  and hence serves as an optimal solution:

$$h(x_1, x_2) = \frac{5}{2}(\frac{20}{19} + \frac{18}{19}t) + (\frac{45}{19} - \frac{45}{19}t) = 5.$$



Finally, we may encounter LP problems with an unbounded feasible region, where the objective function can achieve arbitrarily large values, resulting in the absence of optimal solutions (unbounded objective). For example, this occurs when attempting to maximize  $x_1 + x_2$  in the first quadrant without any additional constraints, or when adding the constraint  $x_2 - x_1 \leq 1$ .

Another possible scenario involves an unbounded linear programming problem that still has infinitely many optimal solutions. This situation is illustrated by the following example, where the details left as an exercise.

**3.A.3.** Maximize  $h = 2x_2 - x_1$  subject to the constraints  $x_1 - x_2 \geq -1, -0.5x_1 + x_2 \leq 2, x_1 \geq 0, x_2 \geq 0$ .

Linear programming applies to a wide range of problems and in the sequel we will explore several of its applications. Examples include maximizing the profits of a production process, minimizing the area on a chip, optimizing

How to solve such a problem? We seek the maximum value of a linear form  $h$  over subsets  $M$  of a vector space which are defined by linear inequalities. In the plane,  $M$  is given by the intersection of half planes.

Next, note that every linear form over real vector space  $h : V \rightarrow \mathbb{R}$  (that is, arbitrary linear scalar function) is monotone in every chosen direction. More precisely, if we choose a fixed starting vector  $u \in V$  and “directional” vector  $v \in V$ , then composition of our form  $h$  with parametrization yields

$$t \mapsto h(u + tv) = h(u) + th(v).$$

This expression is indeed either increasing or decreasing, or constant (depending on whether  $h(v)$  is positive, negative or zero), as a function of  $t$ .

Thus, if the set  $M$  is bounded as at our picture above, we easily find the solution by testing the value of  $h$  at the vertices of the boundary polygon. In general, we must expect that problems similar to the one with the painter are either unsatisfiable (if the given set with constraints is empty), or the profit is unbounded (if the constraints allow for unbounded directions in the space and the form  $h$  is non-zero in some of the unbounded directions) or they attain a maximal solution in at least one of the “vertices” of the set  $M$ . Normally the maximum is attained at a single point of  $M$ , but sometimes it is attained on a part of the boundary of the set  $M$ .

Try to choose explicit values for the parameters  $w_1, w_2, b_1, b_2, c_1, c_2$ , draw the above picture for these parameters and find the explicit solution to the problem (if it exists)!

**3.1.2. Terminology.** In general we speak of a *linear programming problem* whenever we seek either the maximum or minimum value of a linear form  $h$  over  $\mathbb{R}^n$  on a set bounded by a system of linear inequalities which we call *linear constraints*. The vector on the right side is then called the *vector of constraints*. The linear form  $h$  is also called the *objective function*.<sup>1</sup> In real practice we meet hundreds or thousands of constraints for dozens of variables.

The *standard maximization problem* is defined by seeking a maximum of the objective function while the restrictive inequalities are  $\leq$  and the variables are non-negative. On the other hand, the *standard minimization problem* is defined by seeking a minimum of the objective function while the restrictive inequalities are  $\geq$  and the variables are non-negative.

It is easy to see that every general linear programming problem can be transformed into a standard one of either types. Aside from sign changes, we can work with a decomposition of the variables that have no sign restriction into a difference of two non-negative ones. Without loss of generality we will work only with the standard maximization problem.



<sup>1</sup>Leonid Kantorovich and Tjalling Koopmans shared the 1975 Nobel prize in economics for their formulations and solution of economical and logistics problems in a similar way during the second world war. But it was George B. Dantzig who independently developed general linear programming formulation in the period 1946-49, motivated by planning problems in US Air Force. Among others, he invented the simplex method algorithm.



supply chain logistics, scheduling tasks efficiently, and determining the best investment portfolio mix.

**3.A.4.** A company manufactures two models A and B of a product. Each piece of model A requires ten labor hours for fabrication and one labor hour for finishing and packing. For model B, each piece requires sixteen labor hours for fabrication and two labor hours for finishing and packing. The profit is €500 for each piece of model A and €850 for each piece of model B. Assuming there are 180 labor hours available monthly for fabrication and 20 labor hours available for finishing and packing, how many pieces of model A and B should be manufactured to maximize the profit?

**Solution.** Let us encode the given data in a table:

	model A	model B	labour hours
fabricate	10 hours/piece	16 hours/piece	180 hours
finish & pack	1 hour/piece	2 hours/piece	20 hours
profit	€500/piece	€850/piece	

Suppose that  $x_1$  is the number of pieces of model A and  $x_2$  is the number of pieces of model B that are manufactured. The profit function is given by  $h(x_1, x_2) = 500x_1 + 850x_2$ , and we want to maximize it subject to the following constraints:  $10x_1 + 16x_2 \leq 180$ ,  $x_1 + 2x_2 \leq 20$ , with  $x_1 \geq 0$  and  $x_2 \geq 0$ . Here, the first inequality represents the constraint on the labour hours of fabricating, the second one pertains the labour hours of finishing and packing, and the last two are natural non-negativity conditions. Equivalently we have the LP problem of maximizing the given profit  $h = h(x_1, x_2)$  with respect to the conditions  $5x_1 + 8x_2 \leq 90$ ,  $x_1 + 2x_2 \leq 20$ , with  $x_i \geq 0$  for  $i = 1, 2$ . Draw a figure yourself to see, that this is a bounded LP problem and hence the maximum of  $h$  will appear in one of the corners of the corresponding feasible region. The vertices of the feasible region are  $(18, 0)$ ,  $(0, 10)$ , and  $(10, 5)$ . Since  $h(18, 0) = 9000$ ,  $h(0, 10) = 8500$ ,  $h(10, 5) = 9250$ , we deduce that the company must produce 10 pieces of model A and 5 pieces of model B, to achieve the maximum profit of €9250.

□

**3.A.5. Minimize the costs of feeding.** A stable in the west of Czech Republic purchases fodder for the winter: hay and oats. The nutritional values of the fodder, along with the daily requirements (portions) per foal, are detailed in the following table.

g per kg	Hay	Oats	Requirements
Dry basis	841	860	$\geq 6300$ g
Digestible nitrogen stuff	53	123	$\geq 1150$ g
Starch	0.348	0.868	$\leq 5.35$ g
Calcium	6	1.6	$\geq 30$ g
Phosphate	2.8	3.5	$\leq 44$ g
Natrium	0.2	1.4	$\approx 7$ g
Costs	€1.80	€1.60	

**3.1.3. Formulation using linear equations.** Finding an optimum is not always as simple as in the previous 2-dimensional case. The problem can contain many variables and constraints and even deciding whether the set  $M$  of the feasible points is non-empty can be a problem.

We do not have the ambition to go into detailed theory here. But we mention at least some ideas which show that the solution can be always found, and then we build an effective algorithm solving the problem in the next paragraphs.

We begin by comparison with systems of linear equations – because we understand those well. We write the equations (1)-(3) in 3.1.1 in the general form:

$$A \cdot x \leq b,$$

where  $x$  is now an  $n$ -dimensional vector,  $b$  is an  $m$ -dimensional vector and  $A$  is the corresponding matrix. By an inequality between vectors we mean individual inequalities between all coordinates. We want to maximize the product  $c \cdot x$  for a given row vector of coefficients of the linear form  $h$  and the feasible values of  $x$ . If we add new auxiliary variables  $x_s$ , one for every equation and add another variable  $z$  for the value of the linear form  $h$ , we can rewrite the whole system as a system of linear equations

$$(1) \quad \begin{pmatrix} 1 & -c & 0 \\ 0 & A & E_m \end{pmatrix} \cdot \begin{pmatrix} z \\ x \\ x_s \end{pmatrix} = \begin{pmatrix} z - c \cdot x \\ A \cdot x + x_s \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}$$

where the matrix is composed of the blocks with  $1 + n + m$  columns and  $1 + m$  rows, with corresponding individual components of the vectors. We call the new variables  $x_s$  the *slack variables*. Moreover, we require non-negativity for all coordinates  $x$  and  $x_s$ . If the given system of equations has a solution, we seek values for the variables  $z$ ,  $x$  and  $x_s$ , such that all  $x$  and  $x_s$  are non-negative and  $z$  is maximized.

Specifically, in our problem of the black&white painter from 3.1.1, the system of linear equations looks like this:

$$\begin{pmatrix} 1 & -c_1 & -c_2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & w_1 & w_2 & 0 & 1 & 0 \\ 0 & b_1 & b_2 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} z \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ L \\ W \\ B \end{pmatrix}$$

In paragraph 4.1.11 on page 328 we will discuss this situation from the viewpoint of affine geometry. Now we just notice that being on the boundary of the set of feasible points  $M$  of the problem is equivalent to having some of the slack variables vanishing. Our algorithm will try to move from one such position to another while increasing  $h$ . But we shall need some conceptual preparation first.

**3.1.4. Duality of linear programming.** Consider the real matrix  $A$  with  $m$  rows and  $n$  columns, vector of constraints  $b$  and row vector  $c$  giving the objective function. From this data we can consider two problems of linear programming for  $x \in \mathbb{R}^n$  and  $y \in \mathbb{R}^m$ .



During each daily meal, each foal requires a minimum of 2 kg of oats. The average cost, including transportation, is €1.80 per kg of hay and €1.60 per kg of oats. Design a daily diet for one foal that minimizes costs. ○

As explained earlier, when the feasible region of a linear programming problem with two variables is non-empty and bounded, the maximum value of the given linear cost function  $h$  will be found at one of the extreme points (corners) of this region, in the direction of the normal vector to the line defined by  $h$ . This principle also holds true in higher dimensions, as we will see below.



**3.A.6. Linear programming on  $\mathbb{R}^n$ .** Consider a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  of the form

$$f(x_1, \dots, x_n) = c_0 + c_1x_1 + \dots + c_nx_n$$

for some constants  $c_0, c_1, \dots, c_n$ . Due to the appearance of the constant term  $c_0$  we see that  $f$  is an affine function.<sup>2</sup> Since  $f$  is affine, the difference of the values of  $f$  at the points  $A = (a_1, \dots, a_n)$  and  $B = A + u = (a_1 + u_1, \dots, a_n + u_n)$  equals to  $f(u) = f(u_1, \dots, u_n) = c_1u_1 + \dots + c_nu_n$ . Notice the latter is the dot product of the vectors  $(c_1, \dots, c_n)^T$  and  $(u_1, \dots, u_n)^T$  on  $\mathbb{R}^n$ . Next we will refer to  $f$  by the term “objective function”.

The relation between the scalar product and the cosine of the angle between vectors ensures that fixing the value of the given function  $f$  defines a hyperplane in  $\mathbb{R}^n$  with normal the vector  $(c_1, \dots, c_n)^T$ . This hyperplane splits the space  $\mathbb{R}^n$  into two half-spaces. Clearly the given function grows if moving towards one of those half-spaces and declines towards the other one. This is essentially the same principle as we saw when discussing the visibility of segments in dimension 2 (we checked whether the observer is to the left or to the right from the oriented segment, cf. 1.5.12). At the same time, each linear inequality also defines a half-space and we shall learn the properties of intersections of half-spaces in Chapter 4 (they form the so-called simplexes). These observations lead to an algorithm for finding the extremal values of the linear objective function  $f$  on the set of admissible points defined by linear inequalities.

For the sake of simplicity, we shall deal with the “standard problem” of linear programming. This means that we will treat the task of maximizing the linear function

$$h(x_1, \dots, x_n) = \sum_{j=1}^n c_jx_j = c^T x$$

subject to the conditions  $\sum_{j=1}^n a_{ij}x_j \leq b_i$  and  $x_j \geq 0$ , with  $i = 1, \dots, m$ ,  $j = 1, \dots, n$ , respectively. According to the discussion in 3.1.6, our first task is to add the (non-negative) *slack variables*  $x_s$  (also called *slacks*), one for each of the less-than inequalities to reduce them to equalities:

<sup>2</sup>Recall that affine functions are essentially linear functions plus a constant offset.

DUAL PROBLEMS OF LINEAR PROGRAMMING

**Maximization problem:** Maximize  $c \cdot x$  under the conditions  $A \cdot x \leq b$  and  $x \geq 0$ .

**Minimization problem:** Minimize  $y^T \cdot b$  under the conditions  $y^T \cdot A \geq c$  and  $y \geq 0$ .

We say that these two problems are *dual problems of linear programming*. Before deriving further properties of linear programming we need some terminology.

We say that the problem is *solvable* if there is an *admissible vector*  $x$  (or admissible vector  $y^T$ ) which satisfies all constraints. A solvable maximization (minimization) problem is *bounded*, if the objective function is bounded from above (below) over the set of admissible vectors.

**Lemma (Weak duality theorem).** *If  $x \in \mathbb{R}^n$  is an admissible vector for the standard maximization problem, and if  $y \in \mathbb{R}^m$  is an admissible vector for the dual minimization problem, then*

$$c \cdot x \leq y^T \cdot b$$

**PROOF.** It is a simple observation. Since  $x \geq 0$  and  $c \leq y^T \cdot A$ , it follows that  $c \cdot x \leq y^T \cdot A \cdot x$ . But also  $y \geq 0$  and  $A \cdot x \leq b$ , hence

$$c \cdot x \leq y^T \cdot A \cdot x \leq y^T \cdot b,$$

which is what we wanted to prove. □

We see immediately that if both dual problems are solvable, then they must be bounded. Even more interesting is the following corollary, which is directly implied by the inequality in the previous proof.

**Corollary.** *If there exist admissible vectors  $x$  and  $y$  of dual linear problems such that for the objective functions  $c \cdot x = y^T \cdot b$ , then both are optimal solutions for the corresponding problems.*

**3.1.5. Theorem (Strong duality theorem).** *If a standard problem of linear programming is solvable and bounded, then its dual is also bounded and solvable. There exists an optimal solution for each of the problems, and the optimal values of the corresponding objective functions are equal.*

**PROOF.** As already proved in the latter corollary, once it is established that the values of the objective functions for the dual problems equal, we have the required optimal solutions to both problems. It remains to prove the other implication, i.e. the existence of an optimal solution under the assumptions in the theorem, as well as the fact, that the objective functions share their values in such a case. This will be verified by delivering an efficient algorithm in the next paragraph. □

We notice yet another corollary of the just formulated duality theorem:

**Corollary (Equilibrium theorem).** *Consider two admissible vectors  $x$  and  $y$  for the standard maximization problem and its dual problem as defined in 3.1.4. Then both vectors are*

$\sum_{j=1}^n a_{ij}x_j + (x_s)_i = b_i$ , for all  $i = 1, \dots, m$ . Therefore, our goal is to maximize  $h$  over a vector space of solutions defined by a system of linear equations, while ensuring all coordinate values are non-negative. This represents the “canonical form” of a linear programming problem.

In fact, it is useful to have a notation in which the slacks are more or less indistinguishable from the original variables. Therefore, we will often write  $(x_1, \dots, x_n, x_{n+1}, \dots, x_{n+m})$ , instead of  $(x_1, \dots, x_n, (x_s)_1, \dots, (x_s)_m)$ , and then the above equations take the form

$$\sum_{j=1}^n a_{ij}x_j + x_{n+i} = b_i, \quad i = 1, \dots, m.$$

If there are more general inequalities, we can transform them into our standard form by multiplying them by  $-1$ . Additionally, minimizing  $h$  is equivalent to maximizing  $-h$ . Therefore, we can reduce all linear programming problems to the standard form described above.

*The simplex method is an iterative process where we begin with a less-than optimal “solution” that meets the given equations and non-negativity constraints. We then seek for a new solution that improves upon it by increasing the objective function value. Using the Gauss elimination method, we iterate this process until we reach a solution that cannot be further improved—thus achieving an optimal solution. Next, we will analyze the key steps in detail, summarizing the discussion given in 3.1.1 and 3.1.6. To illustrate these steps clearly, we will demonstrate them using a straightforward example.*



**Guide example.** Consider the task of maximizing  $h(x_1, x_2) = 140x_1 + 100x_2$  under the constraints

$$8x_1 + 8x_2 \leq 960, \quad 4x_1 + 2x_2 \leq 400, \quad 4x_1 + 3x_2 \leq 420,$$

together with the positivity conditions  $x_1 \geq 0, x_2 \geq 0$ . The canonical form occurs by maximizing  $h$  subject to

$$\begin{array}{rcccccc} 8x_1 & + & 8x_2 & + & x_3 & & = & 960, \\ 4x_1 & + & 2x_2 & + & & x_4 & = & 400, \\ 4x_1 & + & 3x_2 & + & & & x_5 & = & 420, \end{array}$$

with  $x_i \geq 0$  for any  $i = 1, \dots, 5$ . The slacks are the variables  $x_3, x_4, x_5$  and in terms of matrices we have

$$\tilde{A} = (A|I_3) = \left( \begin{array}{cc|ccc} 8 & 8 & 1 & 0 & 0 \\ 4 & 2 & 0 & 1 & 0 \\ 4 & 3 & 0 & 0 & 1 \end{array} \right), \quad b = \begin{pmatrix} 960 \\ 400 \\ 420 \end{pmatrix},$$

and  $x = (x_1, x_2, x_3, x_4, x_5)^T$  such that  $\tilde{A}x = b$ .

**3.A.7. The simplex algorithm.** To keep the exposition simple, from now on we restrict ourselves to the case where all  $b_i$  are nonnegative,  $b_i \geq 0$  for any  $i = 1, \dots, m$ .

*Step 1.* Using the canonical form of a LP program, we construct the *initial simplex tableau*. The first row of this tableau consists of the coefficients of  $h$  (with negatives signs), while

*optimal if and only if  $y_i = 0$  for all coordinates with index  $i$  for which  $\sum_{j=1}^n a_{ij}x_j < b_i$  and simultaneously  $x_j = 0$  for all coordinates with index  $j$  such that  $\sum_{i=1}^m y_i a_{ij} > c_j$ .*

**PROOF.** Suppose both relations regarding the zeros among  $x_i$  and  $y_i$  are true.



Since the summands with strict inequality have zero coefficients, we have

$$\sum_{i=1}^m y_i b_i = \sum_{i=1}^m y_i \sum_{j=1}^n a_{ij}x_j = \sum_{i=1}^m \sum_{j=1}^n y_i a_{ij}x_j$$

and for the same reason

$$\sum_{i=1}^m \sum_{j=1}^n y_i a_{ij}x_j = \sum_{j=1}^n c_j x_j.$$

This shows one implication, by the duality theorem.

Suppose now that both  $x$  and  $y$  are optimal vectors. Then

$$\sum_{i=1}^m y_i b_i \geq \sum_{i=1}^m \sum_{j=1}^n y_i a_{ij}x_j \geq \sum_{j=1}^n c_j x_j.$$

But the left- and right-hand sides are equal, and hence there is equality everywhere. If we rewrite the first equality as

$$\sum_{i=1}^m y_i \left( b_i - \sum_{j=1}^n a_{ij}x_j \right) = 0,$$

then we see that it can be satisfied only if the relation from the statement holds. But it is a sum of non-negative numbers and equals zero. From the second equality we similarly derive the second part and the proof is finished.  $\square$

The duality theorem and equilibrium theorem are useful when solving linear programming problems, because they show us relations between zeros among the additional variables and the fulfillment of the constraints. As usual, it is good to know that the problem is solvable in principle and to have some theory related to that, but we still need some clever ideas to make it all into an efficient algorithmic procedure. The next paragraph will provide some insight to this.

**3.1.6. The algorithm.** As already explained, the linear programming problem of maximizing the linear objective function  $h = cx$  under the conditions  $Ax \leq b$  can be turned into solving the system of equations (1) in 3.1.3, where we added the slack variables  $x_s$ . If all entries in  $b$  are non-negative, then the choice of  $x_s = b$  and  $x = 0$  provides an admissible solution of the system with the value of the objective function  $h = 0$ . This is the choice of the origin  $x = 0$  as one of the vertices of the distinguished region  $M$  of the admissible points. We can understand this as choosing the variables  $x_s$  as the *basic variables*, whose values are given by the right hand sides of the equation, while all the other variables are set to zero.



In the general case (allowing for negative entries in  $b$ ), we shall see in 4.1.11 that we always can find an admissible vertex. That is, the choice of the basic variables in the above

the subsequent rows are derived from the matrices  $(A|I)$  and  $b$ , i.e.,

$$\begin{array}{cccccc|c} -c_1 & \cdots & -c_n & 0 & 0 & \cdots & 0 & 0 \\ a_{11} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 & b_1 \\ a_{21} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 & b_2 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & 0 & 0 & \cdots & 1 & b_m \end{array}$$

In this initial tableau, there are  $m$  basic variables and  $n$  non-basic variables, a configuration that remains consistent throughout the process. Recall that a basic variable is one that can take non-zero values, whereas a non-basic variable is one that equals zero in the current solution of the problem. Basic variables correspond to the columns of the tableau, where exactly one entry is equal to one, and all other entries are zero. To initiate the algorithm, we choose all slack variables as basic variables. This is achieved by setting to zero the initial decision variables  $x_1, \dots, x_n$  in the equation  $\sum_{j=1}^n a_{ij}x_j + x_{n+i} = b_i$ , for all  $i$ . This initialization ensures an initial feasible solution, though not necessarily optimal.

For the guide example the initial tableau has the form

$$\begin{array}{c|cccccc|c} & x_1 & x_2 & x_3 & x_4 & x_5 & & \\ \hline R_0 & -h & -140 & -100 & 0 & 0 & 0 & 0 \\ R_1 & x_3 & 8 & 8 & 1 & 0 & 0 & 960 \\ R_2 & x_4 & \textcircled{4} & 2 & 0 & 1 & 0 & 400 \\ R_3 & x_5 & 4 & 3 & 0 & 0 & 1 & 420 \end{array}$$

where  $R_0, \dots, R_3$  denote the rows of the table. An initial feasible solution is given by  $x_1 = x_2 = 0$  and  $x_3 = 960$ ,  $x_4 = 400$ ,  $x_5 = 420$ . This satisfies  $h = 0$  and since we have negative values in  $R_0$ , a better solution exists.

*Step 2.* We now move in the iterated steps (compare the more theoretical explanation in 3.1.6).

(a) We locate the column with the most negative value in the first row of the tableau (locating the first column from the left having a non positive value also works). This column, which we assume here that is the  $j$ th column, is called the *work column*. For the guide example, the work column corresponds to the variable  $x_1$ , since  $-140 < -100$ .

(b) In the work column we pick up the positive entry  $a_{ij}$  in  $A$  which provides the minimal relation  $b_i/a_{ij}$ . If more than one element yields the same smallest ratio, we choose one. We call this entry the *pivot*. For our guide example, the pivot appears in row  $R_2$  circled (this is because  $400/4 < 420/4 < 960/8$ ).

(c) To configure the new tableau we proceed with the Gaussian method. Our goal is to eliminate the current work column, and determine the new pivot and the new basic variable (recall that within each iteration of the simplex method, exactly one variable goes from non-basic to basic and exactly one variable goes from basic to non-basic).

To illustrate this via the guide example, denote by  $\hat{R}_2 = (1, 1/2, 0, 1/4, 0, 100)$  the result of the row  $R_2$  divided by the pivot 4. The elementary row operations  $R_0 \rightarrow R_0 + 140\hat{R}_2$ ,

sense, describing an admissible solution. Next, we shall assume to have such a vertex already.

The idea of the algorithm is to perform equivalent row transformations of the entire system in such a way, that we move to other vertices of the region  $M$  and the function  $h$  increases. In order to move to more interesting vertices in  $M$ , we must bring some of the slack variables to zero while the appropriate column for the unit matrix would move to one of those columns corresponding to the variables  $x$ . A simple check reveals that in order to do this, we must choose some of the negative entries in the first row of the matrix 3.1.3(1), pick up this column and choose a row in such a way that using the Gaussian elimination to push the other entries in this particular column to zero, the right hand sides of the equations remain non-negative. The latter condition means that we have to choose the index  $i$  such that  $b_i/a_{ij}$  is minimal. This entry in the matrix is called the *pivot* for the next step in the elimination. Of course, the non-positive coefficients  $a_{ij}$  are not taken into consideration, since they would not lead to any increase in the objective function. When there are no more negative entries in the first row, we are finished, and the claim is that the optimal value of  $h$  appears in the right hand top corner of the matrix.

Before indicating the proof of all the above claims, we show how all this works for the simple problem from 3.F.1. In practice, the very first column of the matrix in question does not change during the procedure at all, so we can omit it completely. Thus we deal with the matrix tableaux:



$$\begin{array}{cc|cccc|c} -4 & -6 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 2 & 1 & 0 & 0 & 0 & 120 \\ 1 & 4 & 0 & 1 & 0 & 0 & 180 \\ -1 & 1 & 0 & 0 & 1 & 0 & -90 \\ 1 & 0 & 0 & 0 & 0 & 1 & 110 \end{array}$$

We cannot find an admissible solution by fixing  $x_s$  as the basic variables here, since there are negative values in  $b$ . We try to initiate the above algorithm by changing the sign in the last but one row and performing the Gaussian elimination for the very first column aiming to have only the 1 in the last but one row there. We obtain:

$$\begin{array}{cc|cccc|c} 0 & -10 & 0 & 0 & -4 & 0 & 360 \\ \hline 0 & \textcircled{3} & \boxed{1} & 0 & 1 & 0 & 30 \\ 0 & 5 & 0 & \boxed{1} & 1 & 0 & 90 \\ \boxed{1} & -1 & 0 & 0 & -1 & 0 & 90 \\ 0 & 1 & 0 & 0 & 1 & \boxed{1} & 20 \end{array}$$

We choose the boxed entries for the basic variables, this represents the values  $x_1 = 90$ ,  $x_2 = 0$ ,  $x_3 = 30$ ,  $x_4 = 90$ ,  $x_5 = 0$ ,  $x_6 = 20$ , and  $h = 360 = 4 \cdot 90 = -4 \cdot (-90)$  which is an admissible solution. We have also circled the pivot for the next step, i.e. the element in the second column which we want to replace with 1 and eliminate the rest of the column (remember this is the one yielding the smallest ratio with the

$R_1 \rightarrow R_1 - 8\hat{R}_2$ , and  $R_3 \rightarrow R_3 - 4\hat{R}_2$  yield the second simplex tableau, given by

		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	
$\hat{R}_0$	$-h$	0	-30	0	35	0	14000
$\hat{R}_1$	$x_3$	0	4	1	-2	0	160
$\hat{R}_2$	$x_1$	1	1/2	0	1/4	0	100
$\hat{R}_3$	$x_5$	0	1	0	-1	1	20

The variable  $x_4$  is replaced by  $x_1$ , which becomes basic, and hence we moved it to the column of basic variables. Together with  $x_3, x_5$  they are now the new basic variables, and we have the solution  $x_1 = 100, x_2 = 0 = x_4, x_3 = 160, x_5 = 20$  (the values of the basic variables are read from the last column). Since there is still a negative number in  $\hat{R}_0$ , we proceed.

(d) We repeat steps (a), (b) and (c) and terminate the procedure when there are no more given entries in the first row.

For our example by the second tableau one deduces the new work column, that of  $x_2$ , and the new pivot, which is the circled number 1. Applying the elementary row operations  $\hat{R}_0 \rightarrow \hat{R}_0 + 30\hat{R}_3, \hat{R}_1 \rightarrow \hat{R}_1 - 4\hat{R}_3$ , and  $\hat{R}_2 \rightarrow \hat{R}_2 - \frac{1}{2}\hat{R}_3$  we arrive to the third tableau:

		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	
$-h$		0	0	0	5	30	14600
$x_3$		0	0	1	2	-4	80
$x_1$		1	0	0	3/4	-1/2	90
$x_2$		0	1	0	-1	1	20

The new basic variable is  $x_2$  and replaces  $x_5$ . So we have the solution  $x_1 = 90, x_2 = 20, x_3 = 80$  and  $x_4 = 0 = x_5$ . Since the above tableau is the final one, this solution is optimal and we may read the maximal value of  $h$  in the right top corner of the tableau, that is,  $h(90, 20) = 14600$ .

In the solution obtained above all the original variables are among the basic ones and so their values are non-zero. This is based on the assumption  $b \geq 0$ . The problem 3.F.1 presented in the final section of this chapter encodes the more general situation and its explanation via this simplex algorithm is given in 3.1.6.

**3.A.8. A note on duality.** By the theoretical discussion in 3.1.4 we know that when the primal LP problem is a maximization one, its *dual* is a minimization problem and conversely. The number of decision variables in a dual problem is the same as the number of less-than inequalities in the optimization primal problem, while the number of the constrains in the dual problem equals to the number of decision variables in the primal one. Note that when some constraint appears as an equality in the primal (maximization) problem, then the corresponding dual variable is unrestricted. The coefficients of the objective function of the primal problem form the right hand side of the constrains on the dual one, while the right hand side of the primal determines the objective function of the dual LP problem.

last right hand column entry among the positive elements  $-30/3 = 10$  which is less than  $90/5 = 16$  and  $20/1 = 20$ ). This leads to the next admissible vertex in our region  $M$  and, of course the value for  $h$  will increase:

		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$		
		0	0	10/3	0	-2/3	0	460
		0	1	1/3	0	1/3	0	10
		0	0	-5/3	1	-2/3	0	40
		1	0	1/3	0	-2/3	0	100
		0	0	-1/3	0	2/3	1	10

with  $x_1 = 100, x_2 = 10, x_3 = x_5 = 0, x_4 = 40, x_6 = 10$ , and  $h = 460 = 4 \cdot 100 + 6 \cdot 10 = \frac{10}{3} \cdot 120 - \frac{2}{3} \cdot (-90)$ . We still have one of the entries in the first line negative. We circled the next pivot leading to

		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$		
		0	0	9/3	0	0	1	470
		0	1	1/2	0	0	-1/2	5
		0	0	-2	1	0	1	50
		1	0	0	0	0	1	110
		0	0	-1/2	0	1	3/2	15

with the final values  $x_1 = 110, x_2 = 5, x_3 = 0, x_4 = 50, x_5 = 15, x_6 = 0$ , and

$$h = 470 = 4 \cdot 110 + 6 \cdot 5 = \frac{9}{3} \cdot 120 + 1 \cdot 110.$$

Let us remind why we can be sure that this is the optimal solution. Thanks to fact that the first line is exclusively non-negative, we have got admissible solution of the dual problem which leads to the same value as the solution of the original one. Thus the equilibrium theorem claims we are done!



**CORRECTNESS OF THE ALGORITHM.** Let us come back to the above claims in some detail. We should check whether the algorithm provides the right answer when it terminates, but we should also check under which conditions it terminates.

We start with the first part. The special feature of the above algorithm reshuffling the basic variables is that the slack variables parts of the matrix are closely linked to the dual linear programming problem. Moreover, there is an invariant of the entire procedure:

**Claim.** Writing  $(-\hat{c}, \hat{c}_s, \hat{h})$  for the current first line in the matrix and  $(\hat{x}, \hat{x}_s)$  for the current values of the variables, we obtain  $c \cdot \hat{x} = \hat{c}_s \cdot b = \hat{h}$  at each step.

In particular at the moment of the termination of the above algorithm, the coefficients  $y = \hat{c}_s$  in the first row represent admissible values of the dual problem (while the values  $\hat{c}$  stay for the slack variables in the dual problem), and the right hand top corner provides the value of the corresponding objective function  $y^T \cdot b$ . Since the two objective functions are equal, we know that the algorithm provides the optimal solution.

For our guide example the dual problem is the task of minimizing  $960y_1 + 400y_2 + 420y_3$  subject to

$$\begin{aligned} 8y_1 + 4y_2 + 4y_3 &\geq 140, \\ 8y_1 + 2y_2 + 3y_3 &\geq 100, \end{aligned}$$

with  $y_i \geq 0$  for all  $i = 1, 2, 3$ .

The final tableau of the primal problem also provides the solution of the dual one. According to the strong duality theorem (see 3.1.5), the minimal value is again 14600, while the corresponding values of the dual variables  $y_1, y_2$  and  $y_3$  are obtained by the first row in the corresponding final tableau, that is,  $y_1 = 0, y_2 = 5, y_3 = 30$ .

**3.A.9.** For the following LP problems, find their standard form and describe the corresponding dual problem.

(a) Maximize  $h = x_1 + 2.5x_2$  subject to

$$2x_1 + 3x_2 \leq 20, \quad x_1 + x_2 \geq -1, \quad x_1 - 2x_2 = 1,$$

with  $x_1 \geq 0, x_2 \geq 0$ .

(b) Minimize  $h = 2x_1 + 3x_2 + 2x_4$  subject to

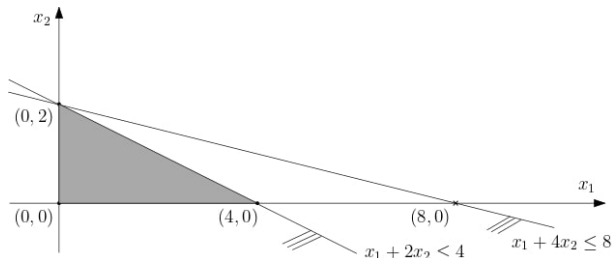
$$x_1 + 2x_2 + 2x_3 \leq -6, \quad x_1 + 4x_2 - 2x_4 = 5, \quad x_2 - x_3 + 4x_4 \geq 2,$$

with  $x_i \geq 0$  for all  $i = 1, \dots, 4$ . ○

**3.A.10. LP problems with redundant constraints.** Use the simplex method to solve the LP problem of maximizing the linear function  $h = 4x_1 + 9x_2$  subject to the conditions

$$x_1 + 4x_2 \leq 8, \quad x_1 + 2x_2 \leq 4, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

**Solution.** The corresponding feasible region is the grey region in the figure below, and as we see the first constraint is redundant. The evaluation of  $h$  at the three corners of this triangle shows that  $h$  maximizes at  $(x_1 = 0, x_2 = 2)$ . We want to verify this result by the simplex method.



The canonical form of the problem is the following: maximize  $h = 4x_1 + 9x_2 + 0x_3 + 0x_4$  subject to

$$\begin{aligned} x_1 + 4x_2 + x_3 &= 8, \\ x_1 + 2x_2 + x_4 &= 4, \end{aligned}$$

with  $x_i \geq 0$  for  $i = 1, \dots, 4$ . The variables  $x_3, x_4$  are the slacks and we begin with the first simplex tableau:

		$x_1$	$x_2$	$x_3$	$x_4$	
$R_0$	$-h$	-4	-9	0	0	0
$R_1$	$x_3$	1	4	1	0	8
$R_2$	$x_4$	1	2	0	1	4

The work column is the column of  $x_2$  and as pivot one may fix the circled number 4. To eliminate the work column replace  $R_1$  by  $\hat{R}_1 := \frac{1}{4}R_1$  and apply the row operations  $R_0 \rightarrow R_0 + 9\hat{R}_1$  and  $R_2 \rightarrow R_2 - 2\hat{R}_1$ . This enters the variable  $x_2$  to the

**PROOF OF THE CLAIM.** As we know, the Gaussian elimination can be expressed as multiplication by a suitable transition matrix  $T$  from the left. Our “pivoting” corresponds always to a (block wise) matrix

$$T = \begin{pmatrix} 1 & y^T \\ 0 & R \end{pmatrix},$$

with an invertible matrix  $R$ . Thus (dealing with the standard problem)

$$T \cdot \begin{pmatrix} 1 & -c & 0 & 0 \\ 0 & A & E_m & 0 \end{pmatrix} = \begin{pmatrix} 1 & -c + y^T A & y^T & y^T b \\ 0 & RA & R & Rb \end{pmatrix}$$

and the second row corresponds (since  $R$  is invertible) to the equality  $A\hat{x} + \hat{x}_s = b$ . Consequently, we arrive at

$$-\hat{c}\hat{x} = (-c + y^T A)\hat{x} = -c\hat{x} + y^T(b - \hat{x}_s).$$

Now, notice the components  $\hat{x}_s$  vanish whenever the corresponding components of  $y^T \neq 0$ , i.e.,  $y^T \hat{x}_s = 0$ . Similarly,  $\hat{c}\hat{x} = 0$  and so we read from the latter displayed equality  $\hat{c}\hat{x} = y^T b$ , as claimed. □

The termination of the algorithm is a more subtle question. While it terminates nearly always, it might happen that the algorithm cycles without enlarging the objective function. We shall not go into more details here, a simple example is given in ??.

Put a cycling simple example to the practice column, maybe in the end of the chapter. □

**3.1.7. Notes about linear models in economy.** Our simple scheme of the black&white painter from the paragraph 3.1.1 can be used to illustrate one of the typical economical models, the *model of production planning*. The model tries to capture the problem completely, that is, to capture both external and internal relations. The left-hand sides of the equations (1), (2), (3) in 3.1.1, and the objective function  $h(x_1, x_2)$  express various production relations. Depending on the character of the problem, we have on the right-hand sides either exact values (and so we solve equations) or capacity constraints and goal optimization (then we obtain linear programming problems).

Thus in general we can solve the problem of source allocation with supplier constraints and either minimize costs or maximize income.

Among economical models, we can find many modifications. One of them is the problem of *financial planning*, which is connected to the optimization of portfolio. We are setting up a volume of investment into individual investment possibilities with the goal to meet the given constraints for risk factors while maximizing the profit, or dually minimize the risk under the given volume.

Another common model is *marketing application*, for instance allocation of costs for advertisement in various media or placing advertisement into time intervals. Restrictions are in this case determined by budget, target population, etc.

Very common are models of *nutrition*, that is, setting up how much of different kinds of food should be eaten in order to meet total volume of specific components, e.g. minerals and vitamins.

column of basic variables, while  $x_3$  leaves. Thus the second simplex tableau has the form

		$x_1$	$x_2$	$x_3$	$x_4$	
$\hat{R}_0$	$-h$	$-5/4$	$0$	$9/4$	$0$	$18$
$\hat{R}_1$	$x_2$	$1/4$	$1$	$1/4$	$0$	$2$
$\hat{R}_2$	$x_4$	$1/2$	$0$	$-1/2$	$1$	$0$

We observe that in this table the basic variable  $x_4$  takes the zero value, and this means that there exists a redundant constraint. Such a basic feasible solution with at least one basic variable equal to zero is called *degenerate*. Now, the column of  $x_1$  is the new driver and the new pivot is the circled  $1/2$ . We replace  $\hat{R}_2$  by  $\check{R}_2 := 2\hat{R}_2$  and apply the row operations  $\hat{R}_1 \rightarrow \hat{R}_1 - \frac{1}{4}\check{R}_2$  and  $\hat{R}_0 \rightarrow \hat{R}_0 + \frac{5}{4}\check{R}_2$ . Thus  $x_1$  enters the column of basic variables,  $x_4$  leaves and the final tableau is given by

		$x_1$	$x_2$	$x_3$	$x_4$	
$\check{R}_0$	$-h$	$0$	$0$	$1$	$5/2$	$18$
$\check{R}_1$	$x_2$	$0$	$1$	$1/2$	$-1/2$	$2$
$\check{R}_2$	$x_1$	$1$	$0$	$-1$	$2$	$0$

This gives us the optimal solution ( $x_2 = 2, x_1 = 0$ ), which is degenerate, with the optimal value of the objective function being  $h = 18$ .  $\square$

**3.A.11. Cycling.** Observe that during the second iteration above, the value of the given  $h$  did *not* increase. When a sequence of iterations returns to a previously constructed tableau, we say that the simplex method *cycles*. In this case the method enters a loop, without improving the objective function value, see also the end of section 3.1.6 Cycling can occur only in the presence of degeneracy. However, there are many LP problems that are degenerate and do *not* cycle (as above). Examples of LP problems for which the simplex method cycles are in general difficult to construct, and rare in practice. We describe such an example in Section 3.F.5, where also a remedy to this pitfall is presented.

Numerous software packages are available to handle the simplex method and solve linear programming problems via a computer. As it is customary in our book, next we will use Sage to implement the tasks at hand.

**3.A.12. LP treated by Sage.** In Sage, for handling linear programming problems, one can utilize the mixed integer linear programming (MILP) package. Next, we will demonstrate the procedure using our guide example from section 3.A.7. Let us summarize the main steps:

1. We initialize the linear programming algorithm:

```
p = MixedIntegerLinearProgram()
```

2. We introduce the non-negative decision variables:

```
v = p.new_variable(real=True,
nonnegative=True)
```

Problems of linear programming arise with personal tasks, where workers with specific qualifications and other properties are distributed into working shifts. Common are also problems of *merging*, problems of *splitting* and problems of *goods distribution*.

## 2. Difference equations

We have already met difference equations in the first chapter, albeit briefly and of first order only. Now we consider a more general theory for linear equations with constant coefficients. This not only provides very practical tools but also represents a good illustration for the concepts of vector spaces and linear mappings.



### HOMOGENEOUS LINEAR DIFFERENCE EQUATION OF ORDER $k$

**3.2.1. Definition.** A homogeneous linear difference equation (or homogeneous linear recurrence) of order  $k$  is given by the expression

$$a_0x_n + a_1x_{n-1} + \dots + a_kx_{n-k} = 0, \quad a_0 \neq 0, \quad a_k \neq 0,$$

where the coefficients  $a_i$  are scalars, which can possibly depend on  $n$ .

A solution of this equation is a sequence of scalars  $x_i$ , for all  $i \in \mathbb{N}$  (or  $i \in \mathbb{Z}$ ), which satisfy the equation with any  $n$ .

We often understand the sequence in question as a function

$$x_n = f(n) = -\frac{a_1}{a_0}f(n-1) - \dots - \frac{a_k}{a_0}f(n-k).$$

By giving any  $k$  consecutive values  $x_i$  in the sequence, all other values of  $x_i$  are determined uniquely. Indeed, we work over a field of scalars, thus the values  $a_0$  and  $a_k$  are invertible and hence, using the recurrent definition, any  $x_n$  can be computed uniquely from the preceding  $k$  values, and similarly for  $x_{n-k}$ . Induction thus immediately proves that all remaining values are uniquely determined.

The space of all infinite sequences  $x_i$  forms a vector space, where addition and multiplication by scalars work coordinate-wise. The definition immediately implies that a sum of two solutions of a homogeneous linear difference equation or a multiple of a solution is again a solution. Analogously as with homogeneous linear systems we see that the set of all solutions forms a subspace.

Initial conditions on the values  $x_0, \dots, x_{k-1}$  of the solution represent a  $k$ -dimensional vector in  $\mathbb{K}^k$ . The sum of initial conditions determines the sum of the corresponding solutions, similarly for scalar multiples. Note also that substituting zeros and ones into initial  $k$  values immediately yields  $k$  linearly independent solutions of the difference equation. Thus, although the vectors are infinite sequences, the set of all solutions has finite dimension. The dimension equals the order of the equation  $k$ . Moreover, we can easily obtain a basis of all those solutions. Again we speak of the *fundamental*

For the guide example there are two decision variables and to introduce them one should additionally type

```
x1, x2 = v["x1"], v["x2"]
```

3. We set the objective function by typing

```
p.set_objective(140*x1 + 100*x2)
```

4. Next we set the constraints. For the guide example this can be done as follows:

```
p.add_constraint(8*x1 + 8*x2 <= 960)
p.add_constraint(4*x1 + 2*x2 <= 400)
p.add_constraint(4*x1 + 3*x2 <= 420)
```

5. If  $k$  is the maximum of the objective function, let the program know about this by typing

```
k = p.solve()
```

Since  $x_1, x_2$  are the coordinates of the solution, we should also include the code

```
x1, x2 = p.get_values(x1, x2)
```

6. Finally, to get the output we write

```
print("Answer =", round(k, 2))
print("(x1, x2) =", (x1, x2))
```

and Sage prints out the following answer:

```
Answer = 14600.0
(x1, x2) = 90.0 20.0
```

Let us summarize the code all together, without interruptions.

```
p = MixedIntegerLinearProgram()
v = p.new_variable(real=True,
nonnegative=True)
x1, x2 = v["x1"], v["x2"]
p.set_objective(140*x1 + 100*x2)
p.add_constraint(8*x1 + 8*x2 <= 960)
p.add_constraint(4*x1 + 2*x2 <= 400)
p.add_constraint(4*x1 + 3*x2 <= 420)
k = p.solve()
x1, x2 = p.get_values(x1, x2)
print("Answer =", round(k, 2))
print("(x1, x2) =", (x1, x2))
```

Similarly can be treated problems with more decision variables. For instance, if you need three variables give the cell

```
v = p.new_variable(real=True,
nonnegative=True)
x1, x2, x3 = v["x1"], v["x2"], v["x3"]
```



Further tasks in linear programming and applications of Sage are described in section F. It is noteworthy that an elegant application of linear programming involves game theory. In 3.F.6 we shall explore an application in this direction concerning zero-sum games, also known as “matrix games”. In these games, the interests of the two players are directly opposed, meaning one player’s loss is the other player’s gain. The term “zero-sum” indicates that the sum of the payoffs for both players always equals zero, regardless of the game’s outcome. As detailed in

system of solutions and all other solutions are its linear combinations.

As we have just checked, if we choose  $k$  indices  $i, i + 1, \dots, i + k - 1$  in sequence, the homogeneous linear difference equation gives a linear mapping  $\mathbb{K}^k \rightarrow \mathbb{K}^\infty$  of  $k$ -dimensional vectors of initial values into infinitely-dimensional sequences of the same scalars. The independence of such solutions is equivalent to the independence of the initial values – which can be easily checked by a determinant: If we have a  $k$ -tuple of solutions  $(x_n^{[1]}, \dots, x_n^{[k]})$ , it is independent if and only if the following determinant, sometimes called the *Casoratian*, is non-zero for one  $n$

$$C_n = \begin{vmatrix} x_n^{[1]} & \cdots & x_n^{[k]} \\ x_{n+1}^{[1]} & \cdots & x_{n+1}^{[k]} \\ \vdots & \ddots & \vdots \\ x_{n+k-1}^{[1]} & \cdots & x_{n+k-1}^{[k]} \end{vmatrix} \neq 0.$$

Notice the determinant  $C_{n+1}$  is obtained from  $C_n$  by replacing the first row in  $C_n$  by the last row in  $C_{n+1}$  expressed as the linear combination of the rows in  $C_n$  given by the difference equation and then putting the first row to position of the last one. Thus  $C_{n+1} = (-1)^k \frac{a_0}{a_k} C_n$ , with values  $a_0, a_k$  corresponding to  $n+k$ . In particular, linearly independent initial conditions lead to independent  $k$ -dimensional vectors for all consecutive  $k$  components of the solution.

**3.2.2. Recurrences with constant coefficients.** It is difficult to find a universal mechanism for finding a solution (that is, a directly computable expression) of general homogeneous linear difference equations. We shall come back to this problem in the end of chapter 13.

In practical models there are very often equations, where the coefficients are constant. In this case it is possible to guess a suitable form for the solution and indeed to find  $k$  linearly independent solutions. This would then be a complete solution of the problem, since all other solutions would be linear combinations of them.



For simplicity we start with equations of second order. Such recurrences are very often encountered in practical problems, where there are relations based on two previous values. A linear difference equation (recurrence) of second order with constant coefficients is thus a formula

$$(1) \quad f(n+2) = a f(n+1) + b f(n) + c,$$


where  $a, b, c$  are known scalar coefficients.

Consider a population model. We assume that the individuals in a population mature and start breeding two seasons later (that is, they add to the value  $f(n+2)$  by a multiple  $b f(n)$  with positive  $b > 1$ ), while immature individuals at the same time weaken and destroy part of the mature population (that is, the coefficient  $a$  at  $f(n+1)$  is negative). Furthermore, it might be that somebody destroys (uses, eats) a fixed amount  $-c$  of individuals every season.



3.F.6, such problems can be effectively addressed using stochastic matrices.

### B. Recurrence relations

 A recurrence relation for a given sequence defines a relationship among its terms, often useful for modeling recurring patterns or behaviours. Such relations are particularly valuable in scenarios where recursive nature is observed repeatedly. For instance, they serve as a powerful tool for describing growth models.


Without doubt, the simplest recurrence relations are the “homogeneous linear recurrence relations” (with constant coefficients), expressed as:

$$x_{n+k} = \alpha_1 x_{n+k-1} + \alpha_2 x_{n+k-2} + \dots + \alpha_k x_n,$$

where  $\alpha_i$  are complex numbers with  $\alpha_k \neq 0$ . The number  $k$  is called the order of the recurrence relation, see also the Section 3.2.1 for details. When the right-hand side includes a constant term, we get the notion of a “non-homogeneous linear recurrence relations” (with constant coefficients).

We begin with a well-known example of a second-order homogeneous linear recurrence relation, historically used to model populations of rabbits.

#### 3.B.1. Rabbits and the Fibonacci sequence.

 Consider a scenario where, at the start of spring, a stork brings two newborn rabbits, one male and one female, to a meadow. Each female rabbit becomes fertile at two months old and gives birth to a pair of newborns, one male and one female, every month thereafter. Specifically, each female is pregnant for one month before producing new offsprings. How many pairs of rabbits will be present after nine months, assuming no deaths and that none “move in”?

**Solution.** After the first month, there is still only one pair of rabbits, but the female is already pregnant. By the end of the second month, the first offspring are born, resulting in two pairs. Each subsequent month, the number of new pairs matches the number of pregnant females from the previous month. This corresponds to the number of at least one month old pairs, which coincides with the number of pairs that were there two months ago.

Let us denote by  $F_n$  the total number of pairs after  $n$  months. According to the explanation given above, for the first months we compute

$$F_1 = 1, F_2 = 1, F_3 = F_2 + F_1 = 2, F_4 = F_3 + F_2 = 3, F_5 = F_4 + F_3 = 5, F_6 = F_5 + F_4 = 8, \dots$$

Hence  $F_n$  should be equal to the sum of the number of pairs in the previous two months, and this is described by the following homogeneous linear recurrence relation:

$$F_{n+2} = F_{n+1} + F_n, \quad n = 1, 2, \dots$$

A similar situation with  $c = 0$  and both other coefficients positive determines the famous Fibonacci sequence of numbers  $y_0, y_1, \dots$ , where  $y_{n+2} = y_{n+1} + y_n$ , see 3.B.1.

If we have no idea how to solve a mathematical problem, we can always blindly try some known solutions of similar problems. Thus, let us substitute into the equation (1) with coefficient  $c = 0$  a similar solution as with the linear equations from the first chapter (cf. 1.2.1), that is, we try  $f(n) = \lambda^n$  for some scalar  $\lambda$ . By substitution into the equation we obtain

$$\lambda^{n+2} - a\lambda^{n+1} - b\lambda^n = \lambda^n(\lambda^2 - a\lambda - b) = 0.$$

This relation will hold either for  $\lambda = 0$  or for the choice of the values

$$\lambda_1 = \frac{1}{2}(a + \sqrt{a^2 + 4b}), \quad \lambda_2 = \frac{1}{2}(a - \sqrt{a^2 + 4b}).$$

It is easy to see that such solutions work. We just had to choose the scalar  $\lambda$  suitably. But we are not finished, since we want to find a solution for any two initial values  $f(0)$  and  $f(1)$ . So far, we have only found two specific sequences satisfying the given equation (or possibly even only one sequence if  $\lambda_2 = \lambda_1$ ).

As we have already derived for linear recurrences, the sum of two solutions  $f_1(n)$  and  $f_2(n)$  of our equation  $f(n+2) - a f(n+1) - b f(n) = 0$  is again a solution of the same equation. The same holds for scalar multiples of the solution. Our two specific solutions thus generate the more general solutions

$$f(n) = C_1 \lambda_1^n + C_2 \lambda_2^n$$

for arbitrary scalars  $C_1$  and  $C_2$ . For a unique solution of the specific problem with given initial values  $f(0)$  and  $f(1)$ , it remains only to find the corresponding scalars  $C_1$  and  $C_2$ .

**3.2.3. The choice of scalars.** We show how this can work with an example. Consider the problem:

$$(1) \quad y_{n+2} = y_{n+1} + \frac{1}{2}y_n, \quad y_0 = 2, y_1 = 0.$$

Here  $\lambda_{1,2} = \frac{1}{2}(1 \pm \sqrt{3})$  and clearly

$$y_0 = C_1 + C_2 = 2$$

$$y_1 = \frac{1}{2}C_1(1 + \sqrt{3}) + \frac{1}{2}C_2(1 - \sqrt{3})$$

is satisfied for exactly one choice of these constants. Direct calculation yields  $C_1 = 1 - \frac{1}{3}\sqrt{3}$ ,  $C_2 = 1 + \frac{1}{3}\sqrt{3}$  and our problem has unique solution

$$f(n) = (1 - \frac{1}{3}\sqrt{3})\frac{1}{2^n}(1 + \sqrt{3})^n + (1 + \frac{1}{3}\sqrt{3})\frac{1}{2^n}(1 - \sqrt{3})^n.$$

Note that even if the found solution for our equation with rational coefficients and rational initial values looks complicated and is expressed with irrational numbers, we know a priori that the solution itself is again rational. But without this “step aside” into a larger field of scalars, we would not be able to describe the general solution.

This relation, along with the initial conditions  $F_1 = 1$  and  $F_2 = 1$ , uniquely determines the number of pairs of rabbits at the meadow, in individual months.

Observe that for certain  $r$  the function  $r^n$  is a solution of the difference equation (without initial conditions). This  $r$  can be obtained by substitution into the recurrent relation:  $r^{n+2} = r^{n+1} + r^n$  and after dividing by  $r^n$  we obtain  $r^2 = r + 1$ . This is the *characteristic equation* of the given recurrent formula, and the numbers  $\frac{1-\sqrt{5}}{2}$  and  $\frac{1+\sqrt{5}}{2}$  are its roots.<sup>3</sup> Thus, according to the theory in 3.2.4, the solution of the given recurrent relation will be a sequence of the form

$$F_n = ax_n + by_n,$$

with  $x_n := (\frac{1-\sqrt{5}}{2})^n$ ,  $y_n := (\frac{1+\sqrt{5}}{2})^n$ , and  $a, b \in \mathbb{R}$  to be specified. To compute the constants  $a, b$ , we use the initial conditions. Alternatively, we can set  $F_0 = 0$  and compute  $a$  and  $b$  from the equations for  $F_0$  and  $F_1$ . We find  $a = -\frac{1}{\sqrt{5}}$ ,  $b = \frac{1}{\sqrt{5}}$  and hence the solution is given by

$$F_n = \frac{1}{\sqrt{5}}(y_n - x_n) = \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{2^n(\sqrt{5})}.$$

Although the Fibonacci sequence looks like a sequence of irrational numbers, the value of  $F_n$  is actually an integer for any natural  $n$ . Hence, all terms in the Fibonacci sequence are integers, as the term  $F_9$ , which gives the answer.  $\square$

*We proceed with additional exercises related to second-order linear homogeneous difference equations with constant coefficients.*

**3.B.2.** Find the next two terms of the sequence  $(a_n)_{n \geq 0}$  beginning as follows:

$$3, 5, 11, 21, 43, 85, \dots$$

by providing a recursive definition of  $a_n$ . Next solve the corresponding recurrence relation.

**Solution.** Set  $a_0 = 3$  and  $a_1 = 5$ . We see that the terms under question are  $a_6 = 171$  and  $a_7 = 341$ . Indeed, observe that  $a_2 = a_1 + 2a_0$ ,  $a_3 = 2a_2 + a_1$ , and so on. Thus the sequence  $a_n$  satisfies  $a_n = a_{n-1} + 2a_{n-2}$  for  $n = 0, 1, \dots$ , and this is the recurrence formula we are seeking for. The corresponding characteristic equation is the quadratic equation  $r^2 - r - 2 = 0$ , with roots the real numbers 2 and  $-1$ . Thus, the general solution has the form  $a_n = a2^n + b(-1)^n$ , with  $a, b \in \mathbb{R}$  to be computed. By the initial conditions we obtain the system  $\{3 = a + b, 5 = 2a - b\}$ , so  $a = 8/3$ ,  $b = 1/3$  and  $a_n = \frac{8}{3}2^n + \frac{1}{3}(-1)^n$ .  $\square$

<sup>3</sup>The number  $\frac{1+\sqrt{5}}{2}$  is called the *golden ratio* and has fascinated mathematicians since Pythagoras (570 – 495 B.C.).

We will often meet similar phenomena. Moreover, the general solution often allows us to discuss qualitative behaviour of the sequence of numbers  $f(n)$  without direct enumeration of the constants. For example, we may see whether the values approach some fixed value with increasing  $n$  or oscillate in some interval or whether they are unbounded.

**3.2.4. General homogeneous recurrences.** We substitute  $x_n = \lambda^n$  for some (yet unknown) scalar  $\lambda$  into the general homogeneous equation from the definition 3.2.1 (with constant coefficients). For every  $n$  we obtain the condition



$$\lambda^{n-k}(a_0\lambda^k + a_1\lambda^{k-1} \dots + a_k) = 0.$$

This means that either  $\lambda = 0$  or  $\lambda$  is the root of the so-called *characteristic polynomial* in the parentheses. The characteristic polynomial is independent of  $n$ .

Assume that the characteristic polynomial has  $k$  distinct roots  $\lambda_1, \dots, \lambda_k$ . For this purpose, we can extend the field of scalars we are working in, for instance  $\mathbb{Q}$  into  $\mathbb{R}$  or  $\mathbb{C}$ . Of course, if the initial conditions are in the original field then the solutions stay there since the recurrence equation itself does. Each of the roots gives us single possible solution

$$x_n = (\lambda_i)^n.$$

We need  $k$  linearly independent solutions.

Thus we should check the independence by substituting  $k$  values for  $n = 0, \dots, k - 1$  for  $k$  choices of  $\lambda_i$  into the Casoratian (see 3.2.1). Thus we obtain the Vandermonde matrix. It is a good but not entirely trivial exercise to show that for every  $k$  and any  $k$ -tuple of distinct  $\lambda_i$  the determinant of such a matrix non-zero, see 2.E.21 on the page 151. It follows that the chosen solutions are linearly independent.

Thus we have found the fundamental system of solutions of the homogeneous difference equation in the case that all the (possibly complex) roots of its characteristic polynomial are distinct.

Now we suppose  $\lambda$  is a multiple root. We ask whether  $x_n = n\lambda^n$  could be a solution. We arrive at the condition

$$a_0n\lambda^n + \dots + a_k(n - k)\lambda^{n-k} = 0.$$

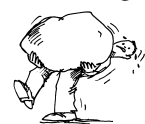
This condition can be rewritten as

$$\lambda(a_0\lambda^n + \dots + a_k\lambda^{n-k})' = 0$$

where the dash denotes differentiation with respect to  $\lambda$  (cf. the infinitesimal definition in 5.1.6, and 12.3.7 for the purely algebraic treatment).

Moreover, a root  $c$  of a polynomial  $f$  has multiplicity greater than one if and only if it is a root of  $f'$ , see 12.3.7 for the proof. Our condition is thus satisfied.

With greater multiplicity  $\ell$  of the root of the characteristic polynomial we can proceed similarly and use the (now obvious) fact that a root with multiplicity  $\ell$  is a root of all derivatives of the polynomial up to order  $\ell - 1$  (inclusively).



**3.B.3.** Find the solution of the recurrence relation  $x_n = 6x_{n-1} - 9x_{n-2}$  with initial conditions  $x_0 = 2$ ,  $x_1 = 3$ . Additionally, validate your answer using Sage.



**Solution.** The characteristic equation has the form

$$r^2 - 6r + 9 = 0 \iff (r - 3)^2 = 0.$$

Hence  $r = 3$  is the unique (double) root. Thus, according to the theory in 3.2.4, the general solution must be of the form

$$x_n = a3^n + bn3^n,$$

for some  $a, b \in \mathbb{R}$  to be computed. Based on the initial conditions  $x_0 = 2$  and  $x_1 = 3$  we obtain  $a = 2$  and  $b = -1$ . Thus we deduce that  $x_n = 2 \cdot 3^n - n3^n = 3^n(2 - n)$ .

Sage provides a user-friendly package for handling recurrence relations. One useful function available is the `rsolve`, from the pure Python package `sympy`, designed for symbolic computations. This function can effectively handle linear recurrence relations. To begin, we typically start by entering the following cell:

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
```

The next step is to input the recurrence relation that we want to solve, along with its initial conditions:

```
f = a(n) - 6*a(n-1) + 9*a(n-2)
initial = {a(0):2, a(1):3}
```

Now we can obtain the solution, by adding the syntax

```
rsolve(f, a(n), initial).expand()
```

Sage prints out the answer  $-3 * n * n + 2 * 3 * n$ . Let us summarize all the code together:

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n) - 6*a(n-1) + 9*a(n-2)
initial = {a(0):2, a(1):3}
rsolve(f, a(n), initial).expand()
```

□

**3.B.4.** Determine an explicit expression of the sequence satisfying the difference equation  $x_{n+2} = 3x_{n+1} + 3x_n$  with members  $x_1 = 1$  and  $x_2 = 3$ . Next verify your answer using Sage.



*Note that characteristic polynomial of a homogeneous difference equation may have complex roots. Therefore, when solving recurrence relations, we may encounter situations involving complex numbers. This requires transitioning from complex bases in the solution space to real ones, as discussed in Section 3.2.5. Let us explore such an example.*

Derivatives look like this:

$$\begin{aligned} f(\lambda) &= a_0\lambda^n + \dots + a_k\lambda^{n-k} \\ f'(\lambda) &= a_0n\lambda^{n-1} + \dots + a_k(n-k)\lambda^{n-k-1} \\ f''(\lambda) &= a_0n(n-1)\lambda^{n-2} + \dots + a_k(n-k)(n-k-1)\lambda^{n-k-2} \\ &\vdots \\ f^{(\ell)} &= a_0n \dots (n-\ell+1)\lambda^{n-\ell} + \dots \\ &\quad + a_k(n-k) \dots (n-k-\ell+1)\lambda^{n-k-\ell}. \end{aligned}$$

We look at the case of a triple root  $\lambda$  and try to find a solution in the form  $n^2\lambda^n$ . By substitution into the definition, we obtain the equation

$$a_0n^2\lambda^n + \dots + a_k(n-k)^2\lambda^{n-k} = 0.$$

Clearly the left side equals the expression  $\lambda^2 f''(\lambda) + \lambda f'(\lambda)$  and because  $\lambda$  is a root of both derivatives, the condition is satisfied.

Using induction, we prove that even for the general condition of the solution in the form  $x_n = n^\ell \lambda^n$ ,

$$a_0n^\ell \lambda^n + \dots + a_k(n-k)^\ell \lambda^{n-k} = 0,$$

the solution can be obtained as a linear combination of the derivatives of the characteristic polynomial starting with the expression (check the combinatorics!)

$$\lambda^\ell f^{(\ell)} + \binom{\ell}{2} \lambda^{\ell-1} f^{(\ell-1)} + \dots$$

We have thus come close to the complete proof of the following result:

HOMOGENOUS EQUATIONS WITH CONSTANT COEFFICIENTS

**Theorem.** *The solution space of a homogeneous linear difference equation of order  $k$  with constant coefficients, over the field of scalars  $\mathbb{K} = \mathbb{C}$ , is the  $k$ -dimensional vector space generated by the sequences  $x_n = n^\ell \lambda^n$ , where  $\lambda$  are (complex) roots of the characteristic polynomial and the powers  $\ell$  run over all natural numbers  $0, \dots, r_\lambda - 1$ , where  $r_\lambda$  is the multiplicity of the root  $\lambda$ .*

**PROOF.** The relation between the multiplicity of roots and the derivatives of real polynomials will be proved later (cf. 5.3.7), while the fact that every complex polynomial has exactly as many roots (counting multiplicities) as its degree will appear in 10.2.11. It remains to prove that the  $k$ -tuple of solutions thus found is linearly independent. Even in this case we can prove inductively that the corresponding Casoratian is non-zero. We have done this already in the case of the Vandermonde determinant before.

This approach is well illustrated by the calculation in the case of a root  $\lambda_1$  with multiplicity one and a root  $\lambda_2$  with

**3.B.5. Complex bases versus real bases.** Determine the solution of the difference equation  $x_{n+2} = 2x_{n+1} - 2x_n$ , with initial values  $x_1 = 2 = x_2$ . Moreover, express your solution in terms of a real basis corresponding to the associated solution space.

**Solution.** The roots of the characteristic polynomial  $r^2 - 2r + 2$  are the complex numbers  $1 + i$  and  $1 - i$ . Hence the sequences  $y_n = (1 + i)^n$  and  $z_n = (1 - i)^n$  form a basis of the (complex) vector space of solutions, see also the main theorem in 3.2.4. In particular, a general solution is given as a linear combination of  $y_n, z_n$  (with complex coefficients), that is,  $x_n = ay_n + bz_n$ , where  $a = a_1 + ia_2, b = b_1 + ib_2$ , with  $a_j, b_j \in \mathbb{R}$  for any  $j = 1, 2$ . The initial conditions will guide us to choose these scalars. They yield the following system of equations

$$1 = a_1 + ia_2 + b_1 + ib_2, \quad 2 = (a_1 + ia_2)(1 + i) + (b_1 + ib_2)(1 - i)$$

and by comparing the real and the complex part of both equations, we finally result to a system of four equations with four unknowns:  $a_1 + b_1 = 1, a_2 + b_2 = 0, a_1 - a_2 + b_1 + b_2 = 2$  and  $a_1 + a_2 - b_1 + b_2 = 0$ . We compute  $a_1 = b_1 = b_2 = \frac{1}{2}, a_2 = -1/2$ , and thus the sequence in question attains the following form:

$$x_n = \left(\frac{1}{2} - \frac{1}{2}i\right)(1 + i)^n + \left(\frac{1}{2} + \frac{1}{2}i\right)(1 - i)^n.$$

If we want to verify our solution via Sage, we may proceed as before, that is,

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n+2) - 2*a(n+1) + 2*a(n)
initial = {a(1):2, a(2):2}
rsolve(f, a(n), initial)
```

Sage's answer has the form

$$\left(\frac{1}{2} + \frac{1}{2}i\right)(1 - i)^n + \left(\frac{1}{2} - \frac{1}{2}i\right)(1 + i)^n$$

For the final task we use the sequences

$$u_n = \frac{1}{2}(y_n + z_n) = (\sqrt{2})^n \cos\left(\frac{n\pi}{4}\right),$$

$$v_n = \frac{i}{2}(z_n - y_n) = (\sqrt{2})^n \sin\left(\frac{n\pi}{4}\right).$$

Note that the transition matrix for changing the basis from the complex one to the real one, is given by  $T := \begin{pmatrix} \frac{1}{2} & -\frac{1}{2}i \\ \frac{1}{2} & \frac{1}{2}i \end{pmatrix}$ , with  $T^{-1} = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}$ . Therefore, if  $(c, d)$  are the coordinates of the sequence  $x_n$ , then under the basis  $\{u_n, v_n\}$  we have

$$\begin{pmatrix} c \\ d \end{pmatrix} = T^{-1} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Thus, we now have an alternative expression for the sequence  $x_n$ , which involves square roots instead of complex numbers:  $x_n = (\sqrt{2})^n \cos\left(\frac{n\pi}{4}\right) + (\sqrt{2})^n \sin\left(\frac{n\pi}{4}\right)$ .  $\square$

multiplicity two:

$$\begin{aligned} C(\lambda_1^n, \lambda_2^n, n\lambda_2^n) &= \begin{vmatrix} \lambda_1^n & \lambda_2^n & n\lambda_2^n \\ \lambda_1^{n+1} & \lambda_2^{n+1} & (n+1)\lambda_2^{n+1} \\ \lambda_1^{n+2} & \lambda_2^{n+2} & (n+2)\lambda_2^{n+2} \end{vmatrix} \\ &= \lambda_1^n \lambda_2^{2n} \begin{vmatrix} 1 & 1 & n \\ \lambda_1 & \lambda_2 & (n+1)\lambda_2 \\ \lambda_1^2 & \lambda_2^2 & (n+2)\lambda_2^2 \end{vmatrix} \\ &= \lambda_1^n \lambda_2^{2n} \begin{vmatrix} 1 & 1 & n \\ \lambda_1 - \lambda_2 & 0 & \lambda_2 \\ \lambda_1(\lambda_1 - \lambda_2) & 0 & \lambda_2^2 \end{vmatrix} \\ &= -\lambda_1^n \lambda_2^{2n} \begin{vmatrix} \lambda_1 - \lambda_2 & \lambda_2 \\ \lambda_1(\lambda_1 - \lambda_2) & \lambda_2^2 \end{vmatrix} \\ &= \lambda_1^n \lambda_2^{2n+1} (\lambda_1 - \lambda_2)^2 \neq 0. \end{aligned}$$

In the general case the proof can be carried on inductively in a similar way.  $\square$

**3.2.5. Real basis of the solutions.** For equations with real coefficients, initial real conditions always lead to real solutions (and similarly with scalars  $\mathbb{Z}$  or  $\mathbb{Q}$ ). However, the corresponding fundamental solutions derived using the above theorem might exist only in the complex domain.

We try therefore to find other generators, which will be more convenient. Because the coefficients of the characteristic polynomial are real, each of its roots is either real or the roots are paired as complex conjugates.

If we describe the solution in polar form as

$$\lambda^n = |\lambda|^n (\cos n\varphi + i \sin n\varphi)$$

$$\bar{\lambda}^n = |\lambda|^n (\cos n\varphi - i \sin n\varphi),$$

we see immediately that their sum and difference leads to two linearly independent solutions

$$x_n = |\lambda|^n \cos n\varphi, \quad y_n = |\lambda|^n \sin n\varphi.$$

Difference equations very often appear as a model of dynamics of some system. A nice topic to think about is the connection between the absolute values of individual roots and the stability of the solution. We will not go into details here, because only in the fifth chapter we will speak of convergence of values to some limit value. There is space for some interesting numerical experiments: for instance with oscillations of suitable population or economical models.

**3.2.6. The non-homogeneous case.** As in the case of systems of linear equations we can obtain all solutions of *non-homogeneous linear difference equations*

$$a_0(n)x_n + a_1(n)x_{n-1} + \cdots + a_k(n)x_{n-k} = b(n),$$

where the coefficients  $a_i$  and  $b$  are scalars which might depend on  $n$ , with  $a_0(n) \neq 0, a_k(n) \neq 0$ . Again, we proceed by finding one solution and adding the complete vector space of dimension  $k$  of solutions to the corresponding homogeneous system. Indeed each such sum yields a solution. Since the difference of two solutions of a non-homogeneous system is

*It is often useful to treat recurrence relations using matrices. While we will describe many such cases in Section C, let us briefly provide an example to highlight this elegant application of matrices. For the reader's convenience, our description includes verifications using Sage, especially for matrix computations.*



**3.B.6. The matrix method via Sage.** Using tools from linear algebra, for  $n \geq 0$  find the solution of the difference equation  $x_{n+2} = 2x_{n+1} + 3x_n$ , with  $x_0 = 0$  and  $x_1 = 1$ .

**Solution.** Set  $p_n = \begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix}$  such that  $p_0 = \begin{pmatrix} x_1 \\ x_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ . Then, for  $n \geq 0$  the difference equation under question can be expressed in terms of matrices as

$$\begin{aligned} p_{n+1} &= \begin{pmatrix} x_{n+2} \\ x_{n+1} \end{pmatrix} = \begin{pmatrix} 2x_{n+1} + 3x_n \\ x_{n+1} \end{pmatrix} \\ &= \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = Ap_n, \end{aligned}$$

where  $A = \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix}$ . The matrix  $A$  is usually referred to as the “companion matrix” of the recurrence and satisfies the relations

$$p_1 = Ap_0, \quad p_2 = Ap_1 = A^2p_0, \quad p_3 = Ap_2 = A^3p_0$$

and so forth. For example, we obtain  $p_1 = (2, 1)^T$  and  $p_2 = (7, 2)^T$ , which implies that  $x_2 = 2$ ,  $x_3 = 7$ , and so on. This approach inductively gives us the expression  $p_n = A^n p_0$ . Therefore, to find the vector  $p_n$  and consequently the desired sequence  $x_n$ , we simply need to compute  $A^n$ . This process is straightforward when  $A$  is diagonalizable.

For our case, using Sage and the syntax

```
var("l")
solve(l^2-2*l-3, l)
```

we find that the characteristic polynomial of  $A$ , represented by  $\lambda^2 - 2\lambda - 3$ , has roots  $\lambda_1 = 3$  and  $\lambda_2 = -1$ . Hence  $A$  has two eigenvalues, both with multiplicity one, and as we will see it is diagonalizable. The corresponding eigenvectors, that is, solutions of the equation  $AX = \lambda X$ , are given by  $X_1 = (1, 1/3)^T$  and  $X_2 = (1, -1)^T$ . In order to obtain these expressions we used Sage, via the block

```
A = matrix(QQ, [[2, 3], [1, 0]])
A.eigenvectors_right()
```

This prints out the answer  $[(3, [(1, 1/3)], 1), (-1, [(1, -1)], 1)]$ .<sup>4</sup> Recall that one may verify that  $A$  is diagonalizable, directly via the command `A.is_diagonalizable()`, which in our case prints `True`. Therefore, we have

$$A^n = PD^nP^{-1}, \quad D = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}, \quad P = \begin{pmatrix} 3 & 1 \\ 1 & -1 \end{pmatrix}$$

<sup>4</sup>Recall that in the expression  $(3, [(1, 1/3)])$  the first number 3 corresponds to the first eigenvalue, the pair  $(1, 1/3)$  encodes its eigenvector, that is  $X_1 = (1, 1/3)^T$ , and the last number 1 is the multiplicity of the eigenvalue. Similarly for the second component inside the list.

a solution of the homogeneous system, we obtain all solutions in this way.

When we were working with systems of linear equations, it was possible that there was no solution. This is not possible with difference equations. But it is not always easy to find that one particular solution of a non-homogeneous system, particularly if the behaviour of the scalar coefficients in the equation is complicated. Even for linear recurrences with constant coefficients it may not be easy to find a solution if the right-hand side is complicated.

But we can always try to find a solution in a form similar to the right hand side. Consider the case when the corresponding homogeneous system has constant coefficients and  $b(n)$  is a polynomial of degree  $s$ . The solution can then be found in the form of the polynomial

$$x_n = \alpha_0 + \alpha_1 n + \dots + \alpha_s n^s$$

with unknown coefficients  $\alpha_i$ ,  $i = 1, \dots, s$ . By substitution into the difference equation and comparing the coefficients of the individual powers of  $n$  we obtain a system of  $s + 1$  equations for  $s + 1$  variables  $\alpha_i$ . If this system has a solution, then we have found a solution of our original problem. If it has no solution, we can try again with an increase in the degree  $s$  of the polynomial in question.

For instance, the equation  $x_n - x_{n-2} = 2$  cannot have a constant solution, because substitution of the potential solution  $x_n = \alpha_0$  yields the requirement  $\alpha_0 - \alpha_0 = 0 = 2$ . But by setting  $x_n = \alpha_0 + \alpha_1 n$  we obtain a solution  $x_n = \alpha_0 + n$ , with  $\alpha_0$  arbitrary. Thus the general solution of our equation is

$$x_n = C_1 + C_2(-1)^n + n.$$

We use this method, the *method of indeterminate coefficients* for example in 3.F.8.

**3.2.7. Variation of constants.** Another possible way to solve such an equation is the *variation of constants* method. Here we find first a solution



$$y(n) = \sum_{i=1}^k c_i f_i(n)$$

of the homogeneous equation, where we consider the constants  $c_i$  as functions  $c_i(n)$  of the variable  $n$ . Then we look for a particular solution of the given equation in the form

$$y(n) = \sum_{i=1}^k c_i(n) f_i(n).$$

We illustrate the method on second order equations. Suppose that the homogeneous part of the second order non-homogeneous equation

$$x_{n+2} + a_n x_{n+1} + b_n x_n = f_n$$

has  $x_n^{(1)}$  and  $x_n^{(2)}$  as a basis of solutions. We will be looking for a particular solution of the non-homogeneous equation in the form

$$x_n = A_n x_n^{(1)} + B_n x_n^{(2)}$$

where for simplicity we have used the multiple  $3X_1 = (3, 1)^T$  in the first column of  $P$  (this is also an eigenvector corresponding to  $\lambda_1$ ). We compute  $P^{-1} = \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & -3/4 \end{pmatrix}$ , which

one can verify by the relation  $P^{-1}AP = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = D$ .

With all the necessary ingredients in hand, we now obtain

$$\begin{aligned} A^n &= \begin{pmatrix} 3 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}^n \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & -3/4 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3^n & 0 \\ 0 & (-1)^n \end{pmatrix} \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & -3/4 \end{pmatrix} \\ &= \frac{1}{4} \begin{pmatrix} 3^{n+1} + (-1)^n & 3^{n+1} - 3(-1)^n \\ 3^n - (-1)^n & 3^n + 3(-1)^n \end{pmatrix}. \end{aligned}$$

To verify the expression for  $A^n$  one can Sage, by the following block (this relies on the fact that  $A$  is diagonalizable)

```
A = matrix(SR, [[2, 3], [1, 0]])
P = matrix(SR, [[3, 1], [1, -1]])
n = var('n')
D = matrix(SR, [[3**n, 0], [0, (-1)**n]])
An = P * D * P.inverse()
An
```

Returning back to our initial task we get:

$$\begin{aligned} p_n &= \begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = A^n p_0 \\ &= \frac{1}{4} \begin{pmatrix} 3^{n+1} + (-1)^n & 3^{n+1} - 3(-1)^n \\ 3^n - (-1)^n & 3^n + 3(-1)^n \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \frac{1}{4} \begin{pmatrix} 3^{n+1} + (-1)^n \\ 3^n - (-1)^n \end{pmatrix}. \end{aligned}$$

From this expression It follows that  $x_n = \frac{1}{4}(3^n - (-1)^n)$ .  $\square$

**3.B.7.** Verify the solution given in 3.B.6 via Sage.  $\circ$

*Homogeneous recurrence relations of order higher than two, can essentially be treated in the same way. However, the characteristic polynomial in this case will be of a higher degree, making the overall procedure a bit more complicated. With this in mind, we present the following task.*

**3.B.8.** For the difference equation  $x_{n+4} = x_{n+3} + x_{n+1} - x_n$  find a real basis of the corresponding solution space.  $\circ$

**3.B.9.** Determine an explicit formula for the  $n$ th member of the unique solution  $\{x_n\}_{n=1}^{\infty}$  that satisfies  $-x_{n+3} = 2x_{n+2} + 2x_{n+1} + x_n$  with  $x_1 = x_2 = x_3 = 1$ .  $\circ$

*Non-homogeneous difference equations have numerous applications. Next, we will dedicate some space to describe examples of this kind (see also 3.2.6 and the problem 3.F.8 in the final section). Notice in Sage non-homogeneous recurrence relations can be treated exactly via the same method described above for the homogeneous case.*

with some conditions on  $A_n$  and  $B_n$  to be imposed. We have

$$\begin{aligned} x_{n+1} &= A_{n+1}x_{n+1}^{(1)} + B_{n+1}x_{n+1}^{(2)} = A_nx_{n+1}^{(1)} + B_nx_{n+1}^{(2)} + \\ &\quad (A_{n+1} - A_n)x_{n+1}^{(1)} + (B_{n+1} - B_n)x_{n+1}^{(2)} \\ &= A_nx_{n+1}^{(1)} + B_nx_{n+1}^{(2)} + \delta A_nx_{n+1}^{(1)} + \delta B_nx_{n+1}^{(2)}, \end{aligned}$$

where  $\delta A_n = A_{n+1} - A_n$  and  $\delta B_n = B_{n+1} - B_n$ .

In order to be able to use the same  $A_n, B_n$  in the expression for  $x_{n+1}$ , we impose for all  $n$  the condition

$$\delta A_nx_{n+1}^{(1)} + \delta B_nx_{n+1}^{(2)} = 0.$$

Thus, for all  $n$

$$x_{n+1} = A_nx_{n+1}^{(1)} + B_nx_{n+1}^{(2)},$$

and in particular

$$\begin{aligned} x_{n+2} &= A_{n+1}x_{n+2}^{(1)} + B_{n+1}x_{n+2}^{(2)} \\ &= A_nx_{n+2}^{(1)} + B_nx_{n+2}^{(2)} + \delta A_nx_{n+2}^{(1)} + \delta B_nx_{n+2}^{(2)}. \end{aligned}$$

Now,

$$\begin{aligned} f_n &= x_{n+2} + a_nx_{n+1} + b_nx_n \\ &= A_n(x_{n+2}^{(1)} + a_nx_{n+1}^{(1)} + b_nx_n^{(1)}) + B_n(x_{n+2}^{(2)} + \\ &\quad a_nx_{n+1}^{(2)} + b_nx_n^{(2)}) + \delta A_nx_{n+2}^{(1)} + \delta B_nx_{n+2}^{(2)} \\ &= \delta A_nx_{n+2}^{(1)} + \delta B_nx_{n+2}^{(2)} \end{aligned}$$

Hence the variations  $\delta A_n$  and  $\delta B_n$  are subject to the systems

$$\begin{aligned} \delta A_nx_{n+1}^{(1)} + \delta B_nx_{n+1}^{(2)} &= 0 \\ \delta A_nx_{n+2}^{(1)} + \delta B_nx_{n+2}^{(2)} &= f_n \end{aligned}$$

with solutions (compute the inverse matrix e.g. by means of the algebraic adjoint and the determinant)

$$\begin{aligned} \delta A_n &= A_{n+1} - A_n = -\frac{f_nx_{n+1}^{(2)}}{W_{n+1}} \\ \delta B_n &= B_{n+1} - B_n = \frac{f_nx_{n+1}^{(1)}}{W_{n+1}} \end{aligned}$$

where  $W_{n+1}$  is the Wronski determinant

$$W_{n+1} = \det \begin{pmatrix} x_{n+1}^{(1)} & x_{n+1}^{(2)} \\ x_{n+2}^{(1)} & x_{n+2}^{(2)} \end{pmatrix}.$$

It follows that

$$\begin{aligned} A_n - A_0 &= \sum_{j=0}^{n-1} -\frac{f_jx_{j+1}^{(2)}}{W_{j+1}} \\ B_n - B_0 &= \sum_{j=0}^{n-1} \frac{f_jx_{j+1}^{(1)}}{W_{j+1}}. \end{aligned}$$

**3.B.10.** Determine the sequence of real numbers that satisfies the non-homogeneous difference equation  $2x_{n+2} = -x_{n+1} + x_n + 2$  with initial conditions  $x_1 = 2, x_2 = 3$ .

**Solution.** The general solution of the homogeneous equation is of the form  $a(-1)^n + b(1/2)^n$ . A particular solution is the constant 1. The general solution of the non-homogeneous equation without initial conditions is thus

$$a(-1)^n + b\left(\frac{1}{2}\right)^n + 1.$$

Based on the initial conditions we compute  $a = 1$  and  $b = 4$ . Consequently, the solution is given by the sequence  $x_n = (-1)^n + 4\left(\frac{1}{2}\right)^n + 1$ .  $\square$

**3.B.11.** Find a sequence which satisfies the following non-homogeneous difference equation  $x_{n+2} = x_{n+1} + 2x_n + 1$  with the initial conditions  $x_1 = 2, x_2 = 2$ . Next verify your solution via Sage  $\circ$

### C. Models of growth and iterated processes

*Many population models are based on recurrence relations in vector spaces. In these models, the unknown is not a sequence of numbers, but a sequence of vectors, with matrices serving as the coefficients. Thus, matrices play a crucial role in population dynamics and can be used to simulate population growth. Next we will focus on such problems, and moreover explore how various iterated processes, such as discrete Markov chains, can model everyday situations and offer intriguing insights.*

**3.C.1. Alice and Bob together in savings.** We begin with a model of growth of saved money.

Alice and Bob are artists having a band together, and they are saving money for online shopping during the black Friday, to enrich their musical instruments. As the black Friday belongs to the last week of November, they start saving every December. This first month Alice gives 1€ and Bob gives 2€. Every consecutive month each of Alice and Bob gives as many as last month, plus one half of what the other has given the month before. How much money will Alice and Bob be able to spend together the next black Friday, by applying the above plan for exactly one year? Is it true or false that during the last month Alice and Bob will put away essentially the same amount of money?

**Solution.** Let us denote by  $a_n$  the amount of money that Alice saves in the  $n$ th month, and let  $b_n$  be the corresponding amount saved by Bob. In the first month they deposit  $a_1 = 1, b_1 = 2$ . The forthcoming savings can be encoded as follows:

$$a_{n+1} = a_n + \frac{1}{2}b_n, \quad b_{n+1} = b_n + \frac{1}{2}a_n.$$

Setting  $A_0 = B_0 = 0$  we obtain

$$A_n = \sum_{j=0}^{n-1} -\frac{f_j x_{j+1}^{(2)}}{W_{j+1}}$$

$$B_n = \sum_{j=0}^{n-1} \frac{f_j x_{j+1}^{(1)}}{W_{j+1}}.$$

and the acquired general solution of our recurrence equation is

$$x_n = C_1 x_n^{(1)} + C_2 x_n^{(2)} + \left(\sum_{j=0}^{n-1} -\frac{f_j x_{j+1}^{(2)}}{W_{j+1}}\right) x_n^{(1)} + \left(\sum_{j=0}^{n-1} \frac{f_j x_{j+1}^{(1)}}{W_{j+1}}\right) x_n^{(2)}.$$

This method is used to solve the example 3.F.9.

**3.2.8. Linear filters.** Now we consider infinite sequences

$$x = (\dots, x_{-n}, x_{-n+1}, \dots, x_{-1}, x_0, x_1, \dots, x_n, \dots).$$

As in the case of systems of linear equations, we work with an operation  $T$  that maps the sequence  $x$  to the sequence  $z = Tx$  with elements

$$z_n = a_0 x_n + a_1 x_{n-1} + \dots + a_k x_{n-k}.$$

As already noticed, the sequences  $x = (x_n)$  are vectors with respect to coordinate-wise operations, and the vector space of all such sequences is infinitely-dimensional. The operation  $T$  is clearly a linear mapping on this space.

The sequences can be imagined as discrete values of a signal, often captured in very short time units.  $T$  plays the role of a filter that works with the signal. For example, this is how the sampling of an audio signal looks like. We are interested in estimating the properties such a *linear filter* can have.

Signals are often a linear combination of superimposed parts, which are themselves periodical. From our definition it is clear that *periodic sequences*  $x_n$ , that is, sequences satisfying for some fixed natural number  $p$  and all  $n$

$$x_{n+p} = x_n$$

will also have periodic images  $z = Tx$

$$z_{n+p} = a_0 x_{n+p} + a_1 x_{n-1+p} + \dots + a_k x_{n-k+p}$$

$$= a_0 x_n + a_1 x_{n-1} + \dots + a_k x_{n-k} = z_n$$

with the same period  $p$ .

We are interested in which input periodic sequences  $Tx$  remain roughly the same (up to a scalar multiple), and in which  $Tx$  will be suppressed close to zero values. This means, we are looking for the kernel of our linear mapping  $T$  and other eigen vectors. The kernel is the subspace of sequences given by the homogeneous difference equation

$$a_0 x_n + a_1 x_{n-1} + \dots + a_k x_{n-k} = 0, \quad a_0 \neq 0 \quad a_k \neq 0,$$

which we are able to solve.

Setting  $p_n := a_n + b_n$  for the common savings during the  $n$ th month, we obtain  $p_{n+1} = \frac{3}{2}p_n$ . This is obviously a geometric sequence and hence

$$p_n = 3 \left(\frac{3}{2}\right)^{n-1} .^5$$

Now, the saving period lasts exactly twelve months, so the sum  $p_1 + p_2 + \dots + p_{12}$  represents the total common savings. We compute

$$3 \left( 1 + \frac{3}{2} + \dots + \left(\frac{3}{2}\right)^{11} \right) = 3 \frac{\left(\frac{3}{2}\right)^{12} - 1}{\frac{3}{2} - 1} \approx 772,5.$$

To confirm this computation use the command `sum` in Sage, as follows:

```
var("n"); pn=3*(3/2)**(n-1)
N(sum(pn, n, 1, 12))
```

In this cell we first introduced  $n$  as a symbolic variable, and next programmed Sage to return the numerical approximation of the desired sum. This answers the first question: The next black Friday Alice and Bob will be able to spend almost 773€. To answer the second task, observe that

$$a_{n+1} = a_n + \frac{1}{2}b_n = \frac{1}{2}(a_n + p_n) = \frac{1}{2}a_n + \left(\frac{3}{2}\right)^n .$$

For a quick solution of this recurrence relation we apply Sage.

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n+1) - (1/2)*a(n) - (3/2)**n
initial = {a(1):1, a(2):2}
rsolve(f, a(n), initial)
```

which gives the answer  $(\frac{3}{2})^{*n} - 1/2^{*n}$ . This means that  $a_n = (\frac{3}{2})^n - (\frac{1}{2})^n$ , and hence during the 12th month the economies of Alice should reach the level of 130€, as it occurs by the expression  $(\frac{3}{2})^{12} - (\frac{1}{2})^{12}$ . Bob will save basically the same amount and the statement is true.  $\square$

**3.C.2. Remark.** Obviously, the example can be adapted to higher initial amounts, although the last months Alice and Bod may face problems to follow their plan. For instance, starting with  $a_1 = 10\text{€}$  and  $b_1 = 20\text{€}$ , we get  $p_1 = 30$ . Therefore, in this case one has

$$p_n = 30 \left(\frac{3}{2}\right)^{n-1}, \quad a_{n+1} = \frac{1}{2}a_n + 15 \left(\frac{3}{2}\right)^{n-1},$$

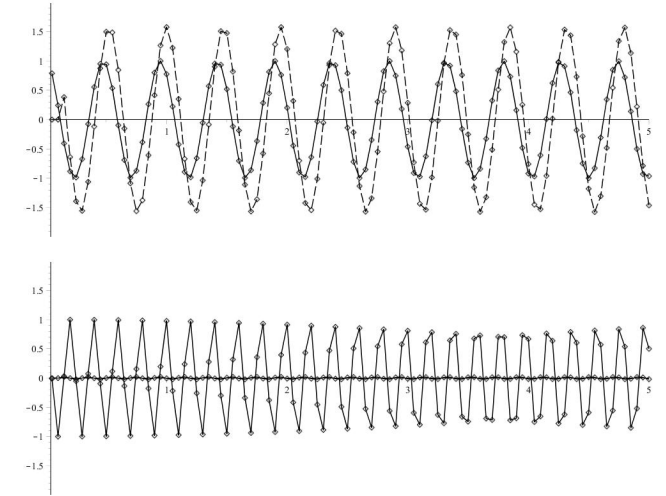
with solution  $a_n = 10\left(\frac{3^n}{2^n} - \frac{1}{2^n}\right)$ . For convenience, we list the relative savings in a table with accuracy two decimal digits.

<sup>5</sup>Recall that the geometric sequence  $x_n = \kappa x_{n-1}$  satisfies  $x_n = x_1 \kappa^{n-1}$ .

**3.2.9. Bad equalizer.** As an example, consider a very simple linear filter given by the equation

$$z_n = (Tx)_n = x_n + x_{n-2}.$$

Clearly, the kernel of  $T$  is generated by  $x_n = \cos(\frac{\pi}{2}n)$  and  $x_n = \sin(\frac{\pi}{2}n)$ , while the solutions to  $x_{n+2} = x_n$  correspond to the requirement  $(Tx)_n = 2x_n$ . The results of such an operation on a signal are illustrated by the two diagrams below. There we use two different frequencies of signals and display their discrete sampling (the solid lines and the points  $x_n$  on them). The dashed line represents the sampling  $z_n$  of the filtered signal.



The first case shows an amplifying of the signal, while the second frequency is close to the kernel which is killed by the filter. Notice that the filtered signal suffers serious shifts in phase, which varies with the frequencies. Cheap equalisers work in such a bad way.

Notice also how badly the original signal is sampled on the second picture. This is due to the fact that the sampling frequency is not much higher than the frequency of the signal.

### 3. Iterated linear processes

**3.3.1. Iterated processes.** In practical models we often encounter the situation where the evolution of a system in a given time interval is given by a linear process, and we are interested in the behaviour of the system after many iterations. The linear process often remains the same, thus from the mathematical point of view we are dealing with an iterated multiplication of the state vector by the same matrix.

While solving the systems of linear equations requires only minimal knowledge of properties of linear mappings, in order to understand the behaviour of an iterated system, we shall exploit the features of eigenvalues, eigenvectors and other structural features.

In fact, the determination of the solution of a linear recurrence equation by a set of initial conditions can be described as an iterated process. Imagine we keep the state vector of the



month	Alice	Bob	month	Alice	Bob
1st	10.00€	20.00€	7th	171.03€	171.43€
2nd	20.00€	25.00€	8th	256.75€	256.95€
3rd	32.50€	35.00€	9th	385.22€	385.32€
4th	50.00€	51.25€	10th	577.89€	577.94€
5th	75.62€	76.25€	11th	866.86€	866.88€
6th	113.75€	114.56€	12th	1300.30€	1300.31€

The dominant chord of the next series of problems is the description of suitable frameworks investigating age-structured population dynamics systems from a linear perspective based on discrete variables. In particular, first we will focus on the “Leslie growth model”, which effectively describes the growth of age-structured populations and is therefore very popular in population ecology.

Recall that this model is written as  $p_{n+1} = Ap_n$ , where  $A$  is the Leslie matrix and  $p_n = (p_n^1, \dots, p_n^m)^T$  is the population vector at time  $n$ , divided into  $m$  age classes. The relevant data for each age class consists of the reproduction rate and the rate of survival into the next age class. These vital rates form the essential part of  $A$ .

The matrix form makes the Leslie model flexible and mathematically very tractable. This is because the solutions of the Leslie model asymptotically exhibit exponential growth. Specifically, the population  $p_n$  grows at an exponential rate determined by  $\lambda_1$ , such that  $p_n \approx \lambda_1^n X_1$ . Here  $\lambda_1$  is the dominant eigenvalue (also known as the “Perron-Frobenius eigenvalue”, see 3.3.3) of  $A$  and  $X_1$  is the corresponding rescaled eigenvector. Furthermore, we can utilize  $X_1$  to infer long-term trends of the age classes, leading to the determination of the stable age distribution. For a more detailed explanation and the proofs of these statements, we encourage the reader to refer to the theoretical section starting at 3.3.2.

**3.C.3. Bushbabies and Leslie matrices.** A small team of biologists studies a colony of bushbabies<sup>6</sup> living in a specific region of South Africa. In the wild, these mammals typically have a lifespan of no more than four years. For research purposes, we can categorize their population into four distinct age classes, as follows:

class A:	0 to 1 year old	class C:	2 to 3 years old
class B:	1 to 2 years old	class D:	3 to 4 years old

The biologists determined that the breeding rates  $b_i$  for the four age classes in this colony are as follows:<sup>7</sup>  $b_1 = 0$ ,  $b_2 = 1$ ,  $b_3 = 2$  and  $b_4 = 1$ , respectively.<sup>8</sup>

<sup>6</sup>Also known as galagos or little night monkeys.

<sup>7</sup>Breeding rates are also called fertility rates and it is natural to assume that they are non-negative real numbers.

<sup>8</sup>Galagos give usually birth to a single offspring at their first pregnancy, and then produce twins in subsequent litters. In particular, they may give birth to two sets of twins a year. (source: <https://animaldiversity.org/>, <https://africageographic.com>)

last  $k$  values

$$Y_n = (x_n, \dots, x_{n-k+1})$$

(filled by the initial condition in the beginning of the process). In the next step we update the state vector

$$Y_{n+1} = (x_{n+1}, x_n, \dots, x_{n-k+2}),$$

where the first entry  $x_{n+1} = a_1x_n + \dots + a_kx_{n-k+1}$  is computed by means of a homogeneous difference equation, while the other entries are just a shift by one position with the last one forgotten. The corresponding square matrix of order  $k$  that satisfies  $Y_{n+1} = A \cdot Y_n$  is as follows:

$$A = \begin{pmatrix} a_1 & a_2 & \dots & a_{k-1} & a_k \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

A while ago, we derived an explicit procedure for the complete formula for the solution of such an iterated process with a special type of matrix. In general, it will not be easy even for very similar systems. A typical case is the study of the dynamics of populations in some biological systems which we discuss below.

The characteristic polynomial  $|A - \lambda E|$  of our matrix is

$$p(\lambda) = (-1)^k (\lambda^k - a_1\lambda^{k-1} - \dots - a_k),$$

as we can check directly or by expanding the last column and employing induction on  $k$ .

Thus, the eigenvalues are exactly the roots  $\lambda$  of the characteristic polynomial of the linear recurrence. We should have expected this, because having a nonzero solution  $x_n = \lambda^n$  to the linear recurrence means that the matrix  $A$  must bring  $(\lambda^k, \dots, \lambda)^T$  to its  $\lambda$ -multiple. Thus every such  $\lambda$  must be eigenvalue of the matrix  $A$ .

**3.3.2. Leslie model for population growth.** Imagine that



we are dealing with some system of individuals (cattle, insects, cell cultures, etc.) divided into  $m$  groups (according to their age, evolution stage, etc.). The state  $X_n$  is thus given by the vector

$$X_n = (u_1, \dots, u_m)^T$$

depending on the time  $t_n$  in which we are observing the system. A linear model of evolution of such system is then given by the matrix  $A$  of dimension  $m$ , which gives the change of the vector  $X_n$  to

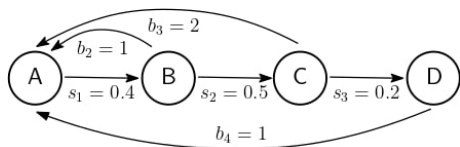
$$X_{n+1} = A \cdot X_n$$

when time changes from  $t_n$  to  $t_{n+1}$ .



Therefore, except for the babies in age class A, all other (female) galagos in the colony can mate and produce offsprings.<sup>9</sup> The biologists also estimated the corresponding survival rates  $s_i$  for these four classes:  $s_1 = 0.4$ ,  $s_2 = 0.5$ ,  $s_3 = 0.2$ , and  $s_4 = 0$ . Notice that

$0 \leq s_i \leq 1$ , as each number represents the probability of a bushbaby surviving from one age class to the next. Graphically, the vital rates and the interactions between the age classes are represented by the following directed graph:



Suppose that the current age distribution of female bushbabies across the four age classes is given by the vector  $p_0 = (20, 20, 30, 10)^T$ . Apply the Leslie model to compute the population of babies and the total female population of the colony after one year. In addition, determine the population of female galagos in age classes C and D ten years later, and deduce that the total female population decreases. Finally provide the long-term trends for the four age classes.

**Solution.** The associated Leslie matrix is given by

$$A = \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & s_3 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 0.4 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \end{pmatrix}.$$

Let us denote by  $A_t, \dots, D_t$  the amounts of the female population belonging to the age class A, ..., D, respectively, at time  $t$ . Then, the Leslie condition  $p_{t+1} = Ap_t$  gives

$$p_{t+1} = \begin{pmatrix} A_{t+1} \\ B_{t+1} \\ C_{t+1} \\ D_{t+1} \end{pmatrix} = \begin{pmatrix} B_t + 2C_t + D_t \\ 0.4A_t \\ 0.5B_t \\ 0.2C_t \end{pmatrix}.$$

Thus,  $p_1 = (A_1, B_1, C_1, D_1)^T = (90, 8, 10, 6)^T$ , and so after one year the model predicts the existence of 90 babies and in total a number of 114 female galagos (this is represented by the sum  $A_1 + \dots + D_1$ ).

For the second task, we utilize the rule  $p_{10} = Ap_9 = A^{10}p_0$ . Using the following cell in Sage

```
A=matrix(SR, [[0, 1, 2, 1],
               [0.4, 0, 0, 0],
               [0, 0.5, 0, 0],
               [0, 0, 0.2, 0]])
show(A**10)
```

<sup>9</sup>Note that our focus is solely on the female population.

As an example, we consider the *Leslie model for population growth*. Here there is the matrix

$$A = \begin{pmatrix} f_1 & f_2 & f_3 & \dots & f_{m-1} & f_m \\ \tau_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \tau_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \tau_3 & \ddots & 0 & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \tau_{m-1} & 0 \end{pmatrix},$$

whose parameters are tied with the evolution of a population divided into  $m$  age groups such that  $f_i$  denotes the relative fertility of the corresponding age group (in the observed time shift from  $N$  individuals in the  $i$ -th group arise new  $f_i N$  ones – that is, they are in the first group), while  $\tau_i$  is the relative mortality in the  $i$ -th group in one time interval. Clearly such a model can be used with any number of age groups.

All coefficients are thus non-negative real numbers and the numbers  $\tau_i$  are between zero and one. Note that when all  $\tau$  are equal one, it is actually a linear recurrence with constant coefficients and thus has either exponential growth/decay (for real roots  $\lambda$  of the characteristic polynomial) or oscillation connected with potential growth/decay (for complex roots).

Before we introduce a more general theory, we consider in more detail this specific model.

Direct computation with the Laplace expansion of the last column yields the characteristic polynomial  $p_m(\lambda)$  of the matrix  $A$  for the model with  $m$  groups:

$$p_m(\lambda) = -\lambda p_{m-1}(\lambda) + (-1)^{m-1} f_m \tau_1 \dots \tau_{m-1}.$$


By induction we derive that this characteristic polynomial is of the form

$$p_m(\lambda) = (-1)^m (\lambda^m - a_1 \lambda^{m-1} - \dots - a_{m-1} \lambda - a_m).$$

The coefficients  $a_1, \dots, a_m$ , are all positive if all parameters  $\tau_i$  and  $f_i$  are positive. In particular,

$$a_m = f_m \tau_1 \dots \tau_{m-1}.$$

Consider the distribution of the roots of the polynomial  $p_m(\lambda)$ . We write the characteristic polynomial in the form

  $p_m(\lambda) = \pm \lambda^m (1 - q(\lambda))$  where  $q(\lambda) = a_1 \lambda^{-1} + \dots + a_m \lambda^{-m}$  is a strictly decreasing non-negative function for  $\lambda > 0$ . For  $\lambda$  positive but very small the value of  $q$  will be arbitrarily large, while for large  $\lambda$ , it will be arbitrarily close to zero. Thus, it seems there exists exactly one positive  $\lambda$  for which  $q(\lambda) = 1$  and thus also  $p_m(\lambda) = 0$ .<sup>2</sup> In other words, for every Leslie matrix (with all the parameters  $f_i$  and  $\tau_i$  positive), there exists exactly one positive real eigenvalue. For actual Leslie models of populations a typical situation is when the only real eigenvalue  $\lambda_1$  is greater or equal to one, while the absolute values of the other eigenvalues are strictly less than one.

<sup>2</sup>Actually, we shall spend a lot of time in chapter 5 to make such considerations precise, involving the convergence and continuity issues.

we obtain

$$A^{10} \approx \begin{pmatrix} 0.195 & 0.465 & 0.457 & 0.204 \\ 0.081 & 0.195 & 0.208 & 0.095 \\ 0.047 & 0.102 & 0.099 & 0.044 \\ 0.008 & 0.023 & 0.023 & 0.010 \end{pmatrix}.$$

Therefore, a further computation shows that

$$p_{10} = (A_{10}, \dots, D_{10})^T \approx (28.98, 12.75, 6.44, 1.44)^T.$$

It follows that  $C_{10} \approx 6.4$  and  $D_{10} \approx 1.4$ . We also deduce that after ten years, the total population of female galagos will be significantly smaller, specifically fewer than 50 individuals. The eventual extinction of the population can be inferred from the dominant eigenvalue of  $A$ . To determine this, one must compute the eigenvalues of  $A$  and compare them accordingly.

To compute the eigenvalues we will use Sage. However, instead of employing the conventional method from Chapter 2 (see ??), we will opt for a more convenient approach by obtaining the eigenvalues numerically. This can be done by using the command `A_num.eigenvalues`. This method applies specifically to matrices in numerical form. After defining the matrix  $A$ , the process unfolds as follows:

```
A_num = A.change_ring(RDF); show(A_num)
eig_numeric = A_num.eigenvalues()
show(eig_numeric)
```

Verify that this block accurately computes the eigenvalues of  $A$  in a highly suitable format. For a more comprehensive response, you can replace the final line with

```
print("Numerical eigenvalues:")
for eigenvalue in eig_numeric:
    print(eigenvalue.n(digits=10))
```

According to Sage there are two real eigenvalues,  $\lambda_1 \approx 0.9347$ , and  $\lambda_2 \approx -0.1121$ , and two complex  $\lambda_{3,4} \approx -0.4112 \pm i0.4607$ . Thus  $\lambda_1$  is the dominant eigenvalue of  $A$  and since  $\lambda_1 < 1$ , the colony will disappear with rate  $p_{t+1}/p_t$  equal to  $\lambda_1 \approx 0.9347$ .

An answer of the final task relies on the eigenvector  $\hat{X}_1$  corresponding to  $\lambda_1$ . Adding in the initial cell the code

```
eigV = A_num.eigenvectors_right()
show(eigV)
```

we see that  $\hat{X}_1$  has (approximately) the expression  $(0.8988, 0.3847, 0.2058, 0.04403)^T$ , but it is more convenient to rewrite this vector as  $\hat{X}_1 \approx (1, 0.427, 0.228, 0.048)^T$ . The number 1.703 approximates the sum of its entries, so the normalized eigenvector associated to  $\lambda_1$  has the form  $X_1 = \frac{1}{1.703} \hat{X}_1 \approx (0.587, 0.250, 0.133, 0.028)^T$ . From this we obtain the long term trends of the four age classes in the population: 0.58% of babies, 0.25% of age class B, while the population of female galagos older than two years old is smaller than 16.2%.  $\square$

**3.C.4. Remarks.** To illustrate the decrease of the population in the previous example, we may present a graph where the population's fading is depicted for a period of fifty years. Included is also the situation for babies' population.

If we begin with any state vector  $X$ , given as a sum of eigenvectors

$$X = X_1 + \dots + X_m$$

with eigenvalues  $\lambda_i$ , then iterations yield

$$A^k \cdot X = \lambda_1^k X_1 + \dots + \lambda_m^k X_m.$$

Thus under the assumption that  $|\lambda_i| < 1$  for all  $i \geq 2$ , all components in the eigensubspaces decrease very fast, except for the component  $\lambda_1^k X_1$ .

The distribution of the population among the age groups is thus very fast approaching the ratios of the components of the eigenvector to the dominant eigenvalue  $\lambda_1$ .

As an example, consider the matrix below where individual coefficients are taken from the model for sheep breeding, that is, the values  $\tau$  contain both natural deaths and activities of breeders.

$$A = \begin{pmatrix} 0 & 0.2 & 0.8 & 0.6 & 0 \\ 0.95 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 \end{pmatrix}.$$

The eigenvalues are approximately

$$1.03, 0, -0.5, -0.27 + 0.74i, -0.27 - 0.74i$$

with absolute values 1.03, 0, 0.5, 0.78, 0.78 and the eigenvector corresponding to the dominant eigenvalue is approximately

$$X^T = (30 \ 27 \ 21 \ 14 \ 8).$$

We have chosen the eigenvector whose coordinates sum to 100, thus it directly gives us the percentage distribution of the population.

Suppose instead that we wish for a constant population, and that one year old sheep are removed for consumption. Then we need ask how to decrease  $\tau_2$  so that the dominant eigenvalue would be one.

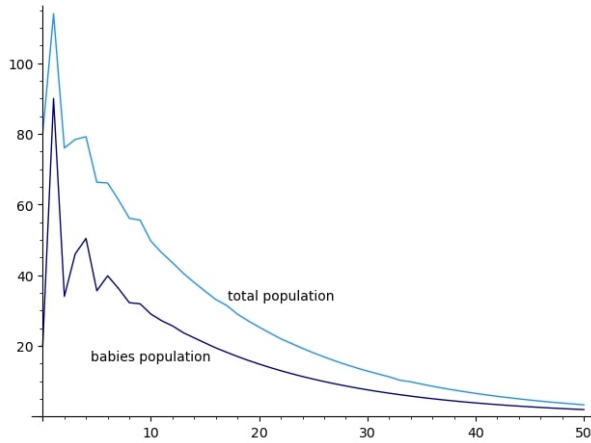
A direct check shows that the farmer could then eat about 10% more of one year old sheep to keep the population constant.

### 3.3.3. Matrices with non-negative elements.

Real matrices which have no negative elements have very special properties. They are very often present in practical models. Thus we introduce the *Perron-Frobenius theory* which deals with such matrices. Actually, we show some results of Perron, we omit the more general situations due to Frobenius.<sup>3</sup>

We begin with some definitions in order to formulate our ideas.

<sup>3</sup>Oskar Perron and Ferdinand Georg Frobenius were two great German mathematicians at the break of the 19th and 20th centuries. Even in this textbook we shall meet their names in Analysis, Number Theory, Algebra. Look up the index.



To obtain the graph, for any plot separately we can apply a command of the following type (we present the code for the first 5 years for the total population and the case of babies population is treated similarly).

```
list_plot({0: 80, 1: 114, 2: 76, 3: 78.4,
4: 79.2, 5: 66.32},
plotjoined=True, color="dodgerblue") \
+ text(r"total population", (22, 34),
color="black")
```

In this block the numbers 0, 1, 2, ... represent the years, and the numbers 80, 114, 76, ... represent the size of the female population the corresponding year. The female population at the  $k$ th year is given by  $m_k = A_k + B_k + C_k + D_k$ . Thus we need to compute the matrices  $A, A^2, \dots, A^{50}$  and then apply the Leslie condition  $p_k = A^k p_0$ , for any  $k$  with  $0 \leq k \leq 50$ . Recall that  $p_k = (A_k, B_k, C_k, D_k)^T$ . Finally, to join the two plots in one figure one can add in the previous cell the code

```
g1 = list_plot(0 : 80, ...), g2 = list_plot(0 : 20, ...),
and then use the command (g1 + g2).show().
```

**3.C.5.** Show that in 3.C.3 the result about the declination of the population will change if we assume that galagos of age class B give two female offsprings, instead of one, and the rest of the given data remains the same. Which are the long term trends of the four age classes in this case and which is the total population of the colony after ten years, if we assume that for any female bushbaby there exists a male one? Present a graph illustrating the increase of the female population for a period of thirty years. ○

*In the study of age-structured populations an interesting question is how we can stabilize the population under examination, so that it is neither growing nor declining. For the Leslie model, this is the case if and only if the dominant eigenvalue of the associated Leslie matrix equals one,  $\lambda_1 = 1$ . Below we describe this situation for two cases: one less realistic, involving the population of female galagos analyzed in 3.C.3, and one more realistic, involving the population of fish in a restricted environment, such as an artificial pond.*



POSITIVE AND PRIMITIVE MATRICES

**Definition.** A *positive matrix* means a square matrix  $A$  all of whose elements  $a_{ij}$  are real and strictly positive. A *primitive matrix* is a square matrix  $A$  whose power  $A^k$  is positive for some positive  $k \in \mathbb{N}$ .

Recall that *spectral radius of a matrix*  $A$  is the maximum of absolute values of all (complex) eigenvalues of  $A$ . The spectral radius of a linear mapping on a (finite dimensional) vector space coincides with the spectral radius of the corresponding matrix for some basis.

In the sequel, the *norm* of a matrix  $A \in \mathbb{R}^{n^2}$  or of a vector  $x \in \mathbb{R}^n$  will mean the sum of the absolute values of all elements. For a vector  $x$  we write  $|x|$  for its norm.

The following result is very useful and hopefully understandable. But the difficulty of its proof is rather not typical for this textbook. If you prefer, read just the theorem and skip the proof till later on.

PERRON THEOREM

**Theorem.** If  $A$  is a primitive matrix with spectral radius  $\lambda \in \mathbb{R}$ , then  $\lambda$  is a root of the characteristic polynomial of  $A$  with multiplicity one and  $\lambda$  is strictly greater than the absolute value of all other eigenvalues of  $A$ . Furthermore, there exists an eigenvector  $x$  associated with  $\lambda$  such that all elements  $x_i$  of  $x$  are positive.

**PROOF.** We shall present the proof only briefly and we shall rely on intuition from elementary geometry, as well as some results touched much later in this book.

Notice that the matrices  $A$  and  $A^k$  share the eigenvectors, while the corresponding eigenvalues are  $\lambda$  and  $\lambda^k$  respectively. Thus the assertion of the theorem holds if and only if the same is true for  $A^k$ . In particular, we may assume the matrix  $A$  itself is positive, without any loss of generality.

Many of the necessary concepts and properties will be discussed in chapter four and in the subsequent chapters devoted to analytical aspects, so the reader might come back to this proof later.

The first step is to show the existence of an eigenvector which has all elements positive. Consider the standard simplex

$$S = \{x = (x_1, \dots, x_n)^T, |x| = 1, x_i \geq 0, i = 1, \dots, n\}.$$

Since all elements in the matrix  $A$  are positive, the image  $A \cdot x$  for  $x \in S$  has all coordinates positive too. The mapping

$$x \mapsto |A \cdot x|^{-1} (A \cdot x)$$

thus maps  $S$  to itself. This mapping  $S \rightarrow S$  satisfies all the assumptions of the *Brouwer fixed point theorem*<sup>4</sup> and thus there

<sup>4</sup>This theorem is a great example of a blend of (homological) Algebra, (differential) Topology and Analysis. We shall discuss it in Chapter 9, cf. 9.1.16 on page 822.

**3.C.6. Stabilizing the size of bushbabies' colony.** Let us return to the colony of bushbabies studied in 3.C.3. Determine the necessary adjustments in the age-dependent vital rates that required to stabilize the population size. By age-dependent vital rates we mean the fertility rates  $b_1, \dots, b_4$  and the survival rates  $s_1, \dots, s_4$ .

**Solution.** Since we are using the Leslie model, the stabilization process is governed by the condition  $\lambda_1 = 1$ , where  $\lambda_1$  is the dominant eigenvalue of the corresponding Leslie matrix. This condition is independent of the initial population size. We consider the following possibilities (though there may be others):

1. Increase  $s := s_1$ , that is the survival rate of babies. Hence we use the Leslie matrix

$$A = \begin{pmatrix} 0 & 1 & 2 & 1 \\ s & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \end{pmatrix}$$

for some  $s$ , with  $0 \leq s \leq 1$ . Since we set  $\lambda_1 = 1$ , to determine the survival rate  $s$  we need to compute the determinant of  $A - E$ , where  $E$  is the identity matrix. For this one can use Sage, as follows:

```
s=var("s")
A=matrix(SR, 4, 4, [0, 1, 2, 1, s, 0, 0,
0, 0, 0.5, 0, 0, 0, 0, 0.2, 0])
E=identity_matrix(4)
(A-E).det()
```

which gives  $-2.1 * s + 1$ . Thus the equation  $\det(A - E) = 0$  has the solution  $s = 10/21 = 0.47619$ . Hence from 0.4 which was initially the breeding parameter  $s = s_1$  we need to increase it to 0.47619 to stabilize the size of the female population of bushbabies (under the assumptions that the other vital rates are as in 3.C.3).

2. Change in a proportional way the fertility rates  $b_i$  for  $i = 2, 3, 4$ . This means that we use the following Leslie matrix

$$A = \begin{pmatrix} 0 & s & 2s & s \\ 0.4 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \end{pmatrix}$$

for some  $s > 0$ , which we should specify via the constraint  $\lambda_1 = 1$ . Using Sage we get  $\det(A - E) = -0.84s + 1$ , and the equation of  $\det(A - E) = 0$  gives  $s = 25/21 \approx 1.19048$ . Therefore, by setting  $b_2 = b_4 = 1.19048$  and  $b_3 = 2.38095$ , while maintaining the survival rates  $s_i$  as in 3.C.3, the population is stabilized.  $\square$

**3.C.7. Population of fishes in an artificial pond.** Suppose we have a simple model of a pond where a population of fish lives (e.g., bleak, vimba, or nase). Assume that 20% of babies survive their second year and, from that age on, they are able to reproduce. It is known that approximately 60% of these young fish survive during the third year, and in the following years, mortality can be ignored. We also assume that the birth rate is three times the number of fish that can reproduce.

exists vector  $y \in S$  such that it is mapped by this mapping to itself. That means that

$$A \cdot y = \lambda y, \quad \lambda = |A \cdot y|$$

and we have found an eigenvector that lies in  $S$ . By assumption,  $A \cdot y$  has got all coordinates positive, thus  $y$  must have the same property. Moreover,  $\lambda > 0$ .

In order to prove the rest of the theorem, we consider the mapping given by the matrix  $A$  in a more suitable basis, where the coordinates of the eigenvector would be  $(\lambda, \dots, \lambda)$ . Moreover, we multiply the mapping by the constant  $\lambda^{-1}$ . Thus we work with the matrix  $B$ ,

$$B = \lambda^{-1}(Y^{-1} \cdot A \cdot Y),$$

where  $Y$  is the diagonal matrix with coordinates  $y_i$  of the above eigenvector  $y$  on its diagonal. Evidently  $B$  is also a positive matrix. By the construction, the vector  $z = (1, \dots, 1)^T$  is its eigenvector with eigenvalue 1, because  $Y \cdot z = y$ .

It remains to prove that  $\mu = 1$  is a simple root of the characteristic polynomial of the matrix  $B$  and that all other roots have absolute value strictly smaller than one. Then the proof of the Perron theorem is finished.

In order to do that we use an auxiliary lemma, which is discussed below. Consider for the moment the matrix  $B$  to define the linear mapping  $\psi$  that maps the row vectors

$$u = (u_1, \dots, u_n) \mapsto u \cdot B = v,$$

that is, using multiplication from the right (i.e.,  $B$  is viewed as the matrix of a linear map on one-forms). Since  $z = (1, \dots, 1)^T$  is an eigenvector of the matrix  $B$  (with eigenvalue 1), the sum of the coordinates of the row vector  $v$

$$u \cdot B \cdot (1, \dots, 1)^T = \sum_{i,j=1}^n u_i b_{ij} = \sum_{i=1}^n u_i = 1,$$

whenever  $u \in S$ . Therefore the simplex  $S$  maps onto itself and thus has in  $S$  a (row) eigenvector  $w$  with eigenvalue one (a fixed point, by the Brouwer theorem again). Because some power  $B^k$  is positive by our assumption, the image of the simplex  $S$  under  $B^k$  lies inside of  $S$ .

We continue with the row vectors. Denote by  $P$  the shift of the simplex  $S$  into the origin by the eigenvector  $w$  we have just found. That is,  $P = -w + S$ . Evidently  $P$  is a set containing the origin and is defined by linear inequalities. Clearly,  $\psi(P) = -w + \psi(S) \subset P$  and, thus, the vector subspace  $V \subset \mathbb{R}^n$  generated by  $P$  is invariant with respect to the action of the matrix  $B$  through multiplication of the row vectors from the right. Moreover, if  $\psi(p) \in P$  sits on the boundary of  $P$ , then  $\psi(p + w)$  is on the boundary of  $S$ . Hence restriction of our mapping to  $V$ , and  $P$  itself satisfy the assumptions of the auxiliary lemma discussed below and thus all its eigenvalues are strictly smaller than one.

Now, the entire space decomposes as the sum  $\mathbb{R}^n = V \oplus \text{span}\{w\}$  of invariant subspaces,  $w$  is the eigenvector with eigenvalue 1, while all eigenvalues of the restriction to  $V$  are strictly smaller in absolute value.

Clearly, such a population would fill the pond very quickly. To maintain a balance, we need to introduce a predator, such as esox. Assume that an esox eats approximately 500 mature fish per year. How many predators of this type should be introduced into the pond to keep the population constant?

**Solution.** There are three age classes: babies, young fish and adult fish. Let  $p$  be the number of babies,  $m$  be the number of young fish and  $r$  be the number of adult fish. In terms of vectors, the state of the population in the next year is given by

$$\begin{pmatrix} p \\ m \\ r \end{pmatrix} \mapsto \begin{pmatrix} 3m + 3r \\ 0.2p \\ 0.6m + yr \end{pmatrix}.$$

The relative mortality of the adult fish caused by the predators equals to the difference  $1 - y$ , and we should specify the unknown  $y$ . This model is successfully described by a *generalized Leslie matrix*, having the form

$$A = \begin{pmatrix} 0 & 3 & 3 \\ 0.2 & 0 & 0 \\ 0 & 0.6 & y \end{pmatrix}.$$

To stabilize the population one of the eigenvalues of this matrix should be equal to 1, and this will give us the desired value of the unknown  $y$ . As before, in Sage we may type

```
y=var("y")
A= matrix(SR, 3, 3, [0, 3, 3, 0.2, 0,
0, 0, 0.6, y])
I3 = matrix(RR, [[1, 0, 0], [0, 1, 0],
[0, 0, 1]])
(A-I3).det ( )
```

This prints out the expression  $0.4 * y - 0.04$  with obvious solution  $y = 0.1$ . This means that in the next year only 10% is allowed to survive and the rest should be eaten by the predators. Let us denote the number of predators by  $x$ . Then, together they eat  $500x$  fish, which according to the previous computation should be  $0.9r$ . Consequently, the ratio of the number of fish to the number of predators is given by  $\frac{r}{x} = \frac{500}{0.9}$ . That is, one esox for (approximately) 556 fish.  $\square$

**3.C.8. Predators and preys.** In a population model, let  $D_k$  be the number of predators and  $K_k$  be the number of preys, in month  $k$ . The relation of these between month  $k$  and month  $k + 1$  is given by one of the following three linear systems:

(a)	$D_{k+1} = 0.6 D_k + 0.5 K_k,$ $K_{k+1} = -0.16 D_k + 1.2 K_k,$
(b)	$D_{k+1} = 0.6 D_k + 0.5 K_k,$ $K_{k+1} = -0.175 D_k + 1.2 K_k,$
(c)	$D_{k+1} = 0.6 D_k + 0.5 K_k,$ $K_{k+1} = -0.135 D_k + 1.2 K_k.$

Analyse the behaviour of this model for large time values.

**Solution.** We can encode all three cases as

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} = T_a \cdot \begin{pmatrix} D_{k-1} \\ K_{k-1} \end{pmatrix}, \quad k \in \mathbb{N},$$

The theorem is nearly proved. We have just to consider the problem that the mapping under question was given by multiplication of the row vectors from the right with the matrix  $B$ , while originally we were interested in the mapping given by the matrix  $B$  and multiplication of the column vectors were from the left. But this is equivalent to the multiplication of the transposed column vectors with the transposed matrix  $B$  in the usual way – from the left. Thus we have proved the claim about eigenvalues for the transpose of  $B$ . But transposing does not change the eigenvalues and so the proof is complete.  $\square$

A bounded polyhedron in  $\mathbb{R}^n$  is a nonempty subset defined by linear inequalities, sitting in some large enough ball. Simplex  $S$  from the proof or any its translation are examples. We shall provide concise explanation of all these concepts in chapter 4.

**Lemma.** Consider any bounded polyhedron  $P \subset \mathbb{R}^n$ , containing a ball around origin  $0 \in \mathbb{R}^n$ . If some iteration of the linear mapping  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  maps  $P$  into its interior (that is  $\psi(P) \subset P$  and the image does not intersect with the boundary), then the spectral radius of the mapping  $\psi$  is strictly less than one.

**PROOF.** Consider the matrix  $A$  of the mapping  $\psi$  in the standard basis. Because the eigenvalues of  $A^k$  are the  $k$ -th powers of the eigenvalues of the matrix  $A$ , we may assume (without loss of generality) that the mapping  $\psi$  already maps  $P$  into  $P$ . Clearly  $\psi$  cannot have any eigenvalue  $\lambda$  with absolute value greater than one. This is easy to see, if the eigenvalue is real. In the complex case, there will be a 2-dimensional invariant plane in which  $\psi$  is a multiplication by  $|\lambda|$  composed with a rotation, thus  $|\lambda| > 1$  is in conflict with the invariance of  $P$  again.

Next, we argue by contradiction and assume that there exists an eigenvalue  $\lambda$  with  $|\lambda| = 1$ . Then there are two possibilities, either  $\lambda^k = 1$  for suitable positive integer  $k$  or there is no such  $k$ .

The image of  $P$  is a closed set (that means that if the points in the image  $\psi(P)$  get arbitrarily close to some point  $y$  in  $\mathbb{R}^n$ , then the point  $y$  is also in the image – this is a general feature of the linear maps on finite dimensional vector spaces). By our assumption, the boundary of  $P$  does not intersect with the image. Thus  $\psi$  cannot have a fixed point on the boundary and there cannot even be any point on the boundary to which some sequence of points in the image would converge.

The first argument excludes that some power of  $\lambda$  is one, because such a fixed point of  $\psi^k$  on the boundary of  $P$  would then exist and thus it would be in the image. In the remaining case there would be a two-dimensional subspace  $W \subset \mathbb{R}^n$  on which the restriction of  $\psi$  acts as a rotation by an irrational angle and thus there exists a point  $y$  in the intersection of  $W$  with the boundary of  $P$ . But then the point  $y$  could be approached arbitrarily close by the points from the set  $\psi^k(y)$  (through all iterations) and thus would have to be in the image too. This leads to a contradiction and thus the lemma is proved.  $\square$

where  $T_a := \begin{pmatrix} 0.6 & 0.5 \\ -a & 1.2 \end{pmatrix}$  and  $a \in \{0.16, 0.175, 0.135\}$ . The coefficient  $a$  represents the average number of preys killed by one predator per month.<sup>10</sup> It follows that

$$\begin{pmatrix} D_k \\ K_k \end{pmatrix} = T_a^k \cdot \begin{pmatrix} D_0 \\ K_0 \end{pmatrix}, \quad k \in \mathbb{N}.$$

Using the powers of the matrix  $T_a$  we can determine the evolution of the populations of predators and prey after a very long time. For such a procedure it is useful to summarize in a table the eigenvalues and the eigenvectors of  $T_a$ :

	$\lambda_1$	$\lambda_2$	$X_1$	$X_2$
$a = 0.16$	1	4/5	$(5, 4)^T$	$(5, 2)^T$
$a = 0.175$	19/20	17/20	$(10, 7)^T$	$(2, 1)^T$
$a = 0.135$	21/20	3/4	$(10, 9)^T$	$(10, 3)^T$

To compute  $T_a^k$  we may simply proceed by hand. As we know, if  $X = (X_1 \ X_2)$  is the matrix formed by the eigenvectors, then the matrix  $T_a^k$  occurs as follows:

$$T_a^k = X \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^k \cdot X^{-1}, \quad k \in \mathbb{N}.$$

For instance, for  $a = 0.16$  we have  $X = \begin{pmatrix} 5 & 5 \\ 4 & 2 \end{pmatrix}$ ,  $X^{-1} = \begin{pmatrix} -1/5 & 1/2 \\ 2/5 & -1/2 \end{pmatrix}$ , and  $\begin{pmatrix} 1 & 0 \\ 0 & 0.8 \end{pmatrix}^k \approx \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$  for large  $k$ .

Thus, for large time values we get  $T_{0.16}^k \approx \begin{pmatrix} -1 & 5/2 \\ -4/5 & 2 \end{pmatrix}$ .

In a similar way, for large  $k$  one computes

$$T_{0.175}^k \approx \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad T_{0.135}^k \approx 1.05^k \begin{pmatrix} -1/2 & 5/3 \\ -9/20 & 3/2 \end{pmatrix}.$$

Knowing  $T_a^k$  we are able to proceed with the systems. These are given by

$$\begin{aligned} \begin{pmatrix} D_k \\ K_k \end{pmatrix} &\approx \frac{1}{10} \begin{pmatrix} 5(-2D_0 + 5K_0) \\ 4(-2D_0 + 5K_0) \end{pmatrix}, \\ \begin{pmatrix} D_k \\ K_k \end{pmatrix} &\approx \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \\ \begin{pmatrix} D_k \\ K_k \end{pmatrix} &\approx \frac{1.05^k}{60} \begin{pmatrix} 10(-3D_0 + 10K_0) \\ 9(-3D_0 + 10K_0) \end{pmatrix}, \end{aligned}$$

for  $a = 0.16$ ,  $a = 0.175$  and  $a = 0.135$ , respectively. A qualitative interpretation of these relations can produce valuable conclusions for the future of the populations. In particular:

- (a) If  $2D_0 < 5K_0$ , the sizes of both populations stabilise on non-zero sizes. if  $2D_0 \geq 5K_0$ , both populations die out.
- (b) Both populations die out.
- (c) For  $3D_0 < 10K_0$  we have a population boom of both kinds. For  $3D_0 \geq 10K_0$  both populations die out.

Note that there are many models that effectively describe predator-prey population dynamics. These models can be either discrete iterative models, or continuous models based on

<sup>10</sup>The value of  $a$  does *not* depend on the size of the population (in fact, for stable populations one can estimate the size of  $a$ ). As a consequence, observe that a small change of  $a$  can lead to different conclusions.

**3.3.4. Simple corollaries.** Once we know the Perron theorem, the following very useful claim has a surprisingly simple proof. It shows how strong is the primitivity assumption of a matrix.



**Corollary.** If  $A = (a_{ij})$  is a primitive matrix and  $x \in \mathbb{R}^n$  is its eigenvector with all coordinates non-negative and eigenvalue  $\lambda$ , then  $\lambda > 0$  is the spectral radius of  $A$ . Moreover,

$$\min_{j \in \{1, \dots, n\}} \sum_{i=1}^n a_{ij} \leq \lambda \leq \max_{j \in \{1, \dots, n\}} \sum_{i=1}^n a_{ij}.$$

**PROOF.** Because  $A$  is primitive, we can choose  $k$  such that  $A^k$  has only positive elements. Then  $A^k \cdot x = \lambda^k x$  is a vector with all coordinates strictly positive. Obviously  $\lambda > 0$ .

According to the Perron theorem, the spectral radius  $\mu$  of  $A$  is an eigenvalue and the associated eigenvectors  $y$  have positive coordinates only. Thus we may choose such an eigenvector with the property that the difference  $x - y$  has only strictly positive coordinates. Then for all large positive integer powers  $m$  we have

$$0 < A^m \cdot (x - y) = \lambda^m x - \mu^m y,$$

but also  $\lambda \leq \mu$ . If  $\mu = \lambda + \alpha$ ,  $\alpha > 0$ , then

$$0 < \lambda^m x - (\lambda + \alpha)^m y < \lambda^m (x - y - m \frac{\alpha}{\lambda} y)$$

which is clearly negative for  $m$  large enough. Hence  $\lambda = \mu$ .

It remains to estimate the spectral radius using the minimum and maximum of sums of individual columns of the matrix. We denote them by  $b_{\min}$  and  $b_{\max}$ . Choose  $x$  to be the eigenvector with the sum of coordinates equal to one and count:

$$\begin{aligned} \sum_{i,j=1}^n a_{ij} x_j &= \sum_{i=1}^n \lambda x_i = \lambda \\ \lambda &= \sum_{j=1}^n \left( \sum_{i=1}^n a_{ij} \right) x_j \leq \sum_{j=1}^n b_{\max} x_j = b_{\max} \\ \lambda &= \sum_{j=1}^n \left( \sum_{i=1}^n a_{ij} \right) x_j \geq \sum_{j=1}^n b_{\min} x_j = b_{\min}. \quad \square \end{aligned}$$

Note that for instance all Leslie matrices from 3.3.2, as soon as all their parameters  $f_i$  and  $\tau_j$  are strictly positive, are primitive. Thus we can apply the just derived results to them. (Compare this with the ad hoc analysis of the roots of the characteristic polynomial from 3.3.2)

**3.3.5. Markov chains.** A very frequent and interesting case of linear processes with only non-negative elements in a matrix is a mathematical model of a system which can be in one of  $m$  states with various probabilities. At a given point of time the system is in state  $i$  with probability  $x_i$ . The transition from the state  $i$  to the state  $j$  happens with probability  $t_{ij}$ .



We can write the process as follows: at time  $n$  the system is described by the *stochastic vector* (we also say *probability vector*)  $x_n = (u_1(n), \dots, u_m(n))^T$ .

differential equations. A prominent example is the Lotka and Volterra model, see 8.3.10. Further exercises related to the Leslie model and population growth are presented in Section F, see for example the tasks 3.F.12, 3.F.13 and 3.F.15.  $\square$

*In the remainder of this section, we delve into the discrete “Markov process”, a topic where matrix calculus intersects with probability theory. For a more theoretical treatment, see 3.3.5 and 3.4.7. We begin with a brief introduction to stochastic processes, a concept that will be extensively analyzed in Chapter 10. Here, we primarily focus on “discrete stochastic processes” (cf. 10.2.14).*

*Recall that a “random variable” in the context of a random experiment with sample space  $\Omega$ , is an  $\mathbb{R}$ -valued function  $X : \Omega \rightarrow \mathbb{R}$  satisfying the property that  $\{s : X(s) \in I\}$  is an event for any interval  $I \subseteq \mathbb{R}$ . For further details see 10.2.10.*

*In simpler terms, a random variable assigns a real number to each possible outcome of a random experiment. A random variable  $X$  that can take at most countably many values is termed discrete. For instance, the number of typographical errors on a page of a book represents a discrete random variable. A stochastic (or random) process is essentially a sequence of random variables  $\{X(t) \equiv X_t\}_{t=0}^\infty$ . Such processes depict the evolution over time of a random phenomenon, thereby encapsulating the concept of probabilistic dynamics. In the following we are only interested in discrete random variables and, consequently, in discrete stochastic processes.*

**3.C.9.** Describe examples of discrete stochastic processes.

**Solution.** There is plethora of such examples. Let  $X_n$  be the number of customers served in a bank, or the number of product sales in a market at the end of the  $n$ th working day. Then  $\{X_n : n = 0, 1, \dots\}$  is a discrete stochastic process. Another example occurs when  $X_n$  represents the number of arrivals at an emergency room between midnight and 8.00 am. Or  $X_n$  can represent the number of books sold in a bookstore, or the number of people who attend a concert, etc.  $\square$

*The values that a discrete random variable  $X_n$  takes are known as the states of the system at step  $n$ . We can denote these states using integers  $0, 1, 2, \dots$ , or symbols such as  $s_0, s_1$ , and so on. Given that  $X_{n-1} = j$  and  $X_n = i$ , the system is said to have made a “transition” from state  $s_j$  to state  $s_i$ . This is known as a chain.*

*“Discrete Markov processes”, also known as “Markov chains”, are stochastic processes with a finite number of possible states, where memory effects are strongly limited. This means that the probability having  $X_n = i$ , depends only on the immediate previous state of the system.*

*Next we will assume time-homogeneity, meaning the transition probabilities  $P(X_n = i | X_{n-1} = j)$  are constant*



This means that all components of the vector  $x$  are real non-negative numbers and their sum equals one. Components give the distribution of the probability of individual possibilities for the state of the system. The distribution of the probabilities at time  $n + 1$  is given via multiplication by the transition matrix  $T = (t_{ij})$ , that is,

$$x_{n+1} = T \cdot x_n.$$

Since we assume that the vector  $x$  captures all possible states of the system and moves again to some of these states with the total probability one, all columns of  $T$  are also given by stochastic vectors. We call such matrices *stochastic matrices*. Note that every stochastic matrix maps every stochastic vector  $x$  to a stochastic vector  $Tx$  again:

$$\sum_{i,j} t_{ij}x_j = \sum_j \left( \sum_i t_{ij} \right) x_j = \sum_j x_j = 1.$$

Such a sequence  $x_{n+1} = Tx_n$  is called a (discrete) *Markov process* and the resulting sequence of vectors  $x_0, x_1, \dots$  is called a *Markov chain*  $x_n$ .

Now we can exploit the Perron-Frobenius theory in its full power. Because the sum of the rows of the matrix is always equal to the vector  $(1, \dots, 1)$ , we see that the matrix  $T - E$  is singular and thus one is an eigenvalue of the matrix  $T$ . Furthermore, if  $T$  is a primitive matrix (for instance, when all elements are non-zero), we know from the corollary 3.3.4 that one is a simple root of the characteristic polynomial and all others have absolute value strictly smaller than one. This leads to:

ERGODIC THEOREM

**Theorem.** *Markov processes with primitive matrices  $T$  satisfy:*

- *there exists a unique eigenvector  $x_\infty$  of the matrix  $T$  with the eigenvalue 1, which is stochastic,*
- *the iterations  $T^k x_0$  approach the vector  $x_\infty$  for any initial stochastic vector  $x_0$ .*

**PROOF.** The first claim follows directly from the positivity of the coordinates of the eigenvector derived in the Perron theorem (notice the dominant eigenvalue comes with multiplicity one).

Next, assume that the algebraic and geometric multiplicities of the eigenvalues of the matrix  $T$  are the same. Then every stochastic vector  $x_0$  can be written (in the complex extension  $\mathbb{C}^n$ ) as a linear combination

$$x_0 = c_1 x_\infty + c_2 y_2 + \dots + c_n y_n,$$

where  $y_2, \dots, y_n$  extend  $x_\infty$  to a basis of the eigenvectors. But then the  $k$ -th iteration gives again a stochastic vector

$$x_k = T^k \cdot x_0 = c_1 x_\infty + \lambda_2^k c_2 y_2 + \dots + \lambda_n^k c_n y_n.$$

Now all eigenvalues  $\lambda_2, \dots, \lambda_n$  are in absolute value strictly smaller than one. So all components of the vector  $x_k$  but the first one approach (in norm) zero. But  $x_k$  is still stochastic,





over time and do not depend on  $n$ . This is a common assumption in many Markov chain models. The transition from state  $s_j$  to  $s_i$ , or simply from  $j$  to  $i$ , is encoded by the transition probability  $t_{ij} = P(X_n = i | X_{n-1} = j)$ , compare with 3.3.5. Thus, the dynamics of a discrete-time homogeneous Markov chain with state space consisting of  $m$  states, is described by the  $m \times m$  matrix  $T = (t_{ij})$ . This is known as the transition matrix of the Markov process, from the  $(n - 1)$ th step to the  $n$ th step. In our context,  $T$  is a column-stochastic matrix, meaning that  $t_{ij} \geq 0$  and  $\sum_i t_{ij} = 1$ . In other words, each column of  $T$  is a stochastic vector, (this has non-negative real numbers as entries, that sum to one, see also 3.3.5).

**3.C.10.** Two-state Markov chains are Markov chains having a state space consisting of two elements, typically denoted by  $S = \{0, 1\}$ . They can be used to model various situations, such as the operational status of everyday machines where the probability of a machine being out of operation the next day is known.



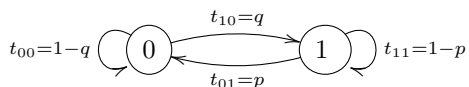
For example, consider a traffic light on a road in New York. Suppose the probability that a working traffic light will be out of order the next day is  $p$ , while the probability that an out-of-order traffic light will start operating the next day is  $q$ . Demonstrate that this scenario constitutes a Markov process and determine its transition matrix.

**Solution.** What happens at each day depends only on the previous day, and hence we have a two-stage (homogeneous) Markov process  $\{X_n\}$ , where  $X_n$  is the state of the traffic light on the  $n$ th day. By definition,  $X_n = 0$  if the traffic light will be out of order the  $n$ th day and  $X_n = 1$  otherwise. Thus one computes

$$\begin{aligned} t_{00} &= P(X_n = 0 | X_{n-1} = 0) = 1 - q, \\ t_{01} &= P(X_n = 0 | X_{n-1} = 1) = p, \\ t_{10} &= P(X_n = 1 | X_{n-1} = 0) = q, \\ t_{11} &= P(X_n = 1 | X_{n-1} = 1) = 1 - p. \end{aligned}$$

This means that  $T = \begin{pmatrix} t_{00} & t_{01} \\ t_{10} & t_{11} \end{pmatrix} = \begin{pmatrix} 1 - q & p \\ q & 1 - p \end{pmatrix}$ .  $\square$

**3.C.11. Remarks.** a) Markov chains on finite state spaces can alternatively be represented graphically through a “transition diagram”. This diagram is a directed graph where each vertex represents a state of the chain. Directed edges are drawn from vertex  $j$  to vertex  $i$ , labeled with the probability  $t_{ij}$ , whenever  $t_{ij} > 0$ . For example, in the case of a two-state Markov process as described earlier, the transition diagram takes the following form:



This graphical representation is particularly useful for visualizing the structure of the Markov chain, understanding the possible state transitions, and analyzing the overall behavior of the chain in terms of state probabilities and sequences of

thus the only possibility is that  $c_1 = 1$  and the second claim is proved.

In fact, even if the algebraic and geometric multiplicities of eigenvalues do not coincide we reach the same conclusion using a more detailed study of the root subspaces of the matrix  $T$ . (We meet them when discussing the Jordan matrix decomposition later in this chapter.) Consequently, even in the general case the eigensubspace  $\text{span}\{x_\infty\}$  comes with the unique invariant  $(n - 1)$ -dimensional complement, on which are all eigenvalues in absolute value smaller than one and the corresponding components in  $x_k$  approach zero as before. See the note 3.4.11 where we finish this argument in detail.  $\square$

**3.3.6. Iteration of the stochastic matrices.** We reformulate the previous theorem into a simple, but surprising result. By convergence to a limit matrix in the following theorem we mean the following: if we say that we want to bound the possible error  $\varepsilon > 0$ , then we can find a lower bound on the number of iterations  $k$  after which all the components of the matrix differ from the limit one by less than  $\varepsilon$ .



**Corollary.** Let  $T$  be a primitive stochastic matrix from a Markov process and let  $x_\infty$  be the stochastic eigenvector for the dominant eigenvalue 1 (as in the Ergodic Theorem above). Then the iterations  $T^k$  converge to the limit matrix  $T_\infty$ , whose columns all equal to  $x_\infty$ .

**PROOF.** Columns in the matrix  $T^k$  are images of the vectors of the standard basis under the corresponding iterated linear mapping. But these are images of the stochastic vectors and thus they all converge to  $x_\infty$ .  $\square$

**3.3.7. Final brief remark.** Before leaving the Markov processes, we shortly mention their more general versions with matrices which are not primitive. Here we would need the full Frobenius-Perron theory. Without going into technicalities, consider a process with a block wise diagonal or an upper triangular matrix  $T$ ,

$$T = \begin{pmatrix} P & R \\ 0 & Q \end{pmatrix}$$

and imagine first that  $P, Q$  are primitive and  $R = 0$ . Here we can again apply the above results block wise. In words, if we start in a stay  $x_0$  with all probability concentrated in the first block of coordinates, the process converges to the value  $x_\infty$  which again has all the probability distributed among the first block of coordinates, and the same for the other block.

If  $R > 0$  then we can always jump to the states corresponding to the first block from those in the second block with a non-zero probability and the iterations get more complicated:

$$\begin{aligned} T^2 &= \begin{pmatrix} P^2 & P \cdot R + R \cdot Q \\ 0 & Q^2 \end{pmatrix} \\ T^3 &= \begin{pmatrix} P^3 & P^2 \cdot R + P \cdot R \cdot Q + R \cdot Q^2 \\ 0 & Q^3 \end{pmatrix}. \end{aligned}$$

transitions, see 3.C.12 for another example.

b) Be aware that other authors may define the transition matrix  $T = (t_{ij})$  with  $t_{ij}$  being the probability of a transition from state  $i$  to state  $j$ , that is  $t_{ij} = P(X_n = j | X_{n-1} = i)$ . This is the opposite convention of ours, producing a row-stochastic matrix whose transpose matrix is our transition matrix.

**3.C.12. Absent-minded professor.** An absent-minded professor always carries an umbrella with him, but with probability of  $1/2$  he forgets it from wherever he is leaving. His daily routine is strict: in the morning he walks from his house to his office; from there, he goes to a restaurant for lunch, then returns to his office, and finally in the evening, he returns home. It is assumed the professor does not visit other locations, and if he forgets his umbrella at the restaurant, it remains there until his next visit. This situation can be modeled as a Markov process. Determine its transition matrix, diagram, and calculate the probability that after many days, in the morning, the umbrella is at the restaurant.

**Solution.** We assume as time unit the daytime, from morning to the next morning, and denote by  $X_n$  the place where the umbrella sits in the morning of the  $n$ th day. This defines a (homogeneous) Markov process  $\{X_n : n \in \mathbb{N}\}$ , with state space  $S = \{s_1 = \text{house}, s_2 = \text{office}, s_3 = \text{restaurant}\}$ . For simplicity we represent these states as  $\{1, 2, 3\}$ . Let us proceed with the transition matrix  $T = (t_{ij})$ , where  $t_{ij} = P(X_{n+1} = i | X_n = j)$  and  $1 \leq i, j \leq 3$ . We explain the computations for the first column and leave the verification of the rest two to the reader. By definition,  $t_{11} = P(X_{n+1} = 1 | X_n = 1)$ , and hence this is the probability that the umbrella starts its day at home and stays there till next morning. There are three distinct possibilities  $p_1, p_2, p_3$  for this scenario:

$p_1$ : The umbrella stays at home in the morning. Thus  $p_1 = \frac{1}{2}$ .  
 $p_2$ : The umbrella arrives to the office, it stays there during lunch, but in the evening is moved back to home. Thus  $p_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ .  
 $p_3$ : The professor takes the umbrella with him all the time and does not forget it somewhere. Thus  $p_3 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$  and in total  $t_{11} = p_1 + p_2 + p_3 = \frac{11}{16}$ .

Next,  $t_{21} = P(X_{n+1} = 2 | X_n = 1)$  is the probability that the umbrella starts its day at home and the next morning sits in the office. There are two possibilities  $q_1, q_2$  for such a scenario:

$q_1$ : The umbrella arrives to the office, it stays there during lunch, and in the evening it remains again there. Thus  $q_1 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ .  
 $q_2$ : The umbrella arrives to the office, then to the restaurant, then back to the office, and in the evening it remains at the office. Thus  $q_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$  and in total  $t_{21} = \frac{1}{8} + \frac{1}{16} = \frac{3}{16}$ . Finally for  $t_{31}$  we have  $t_{31} = 1 - t_{11} - t_{21} = 1/8$ . The repetition of this procedure for the rest entries of  $T$  yields

$$T = \begin{pmatrix} 11/16 & 3/8 & 1/4 \\ 3/16 & 3/8 & 1/4 \\ 1/8 & 1/4 & 1/2 \end{pmatrix}.$$

An interesting special case is when  $P = E$  and  $R$  is positive. Then  $Q - E$  must be a regular matrix and a simple computation yields the general iteration (notice  $E$  and  $Q$  commute and thus  $(E - Q)(E + Q + \dots + Q^{k-1}) = E - Q^k$ )

$$T^k = \begin{pmatrix} E & R(E - Q)^{-1}(E - Q^k) \\ 0 & Q^k \end{pmatrix}.$$

Thus, the entire first block of states is formed by eigenvectors with eigenvalue 1 (so these states stay constant with probability 1), while the behavior on the other block is more complicated.

### 4. More matrix calculus

We have seen that understanding the inner structure of matrices is a strong tool for both computation and analysis. It is even more true when considering numerical calculations with matrices. Therefore we return now to the abstract theory.

We introduce special types of linear mappings on vector spaces. We consider general linear mappings whose structure is understood in terms of the Jordan normal form (see 3.4.10). In all these cases, complex scalars are essential. So we extend our discussion of scalar product (see 2.3.18–2.3.22) to complex vector spaces. Actually, in many areas the complex vector spaces are the essential platform necessary for introducing the mathematical models. For instance, this is the case in the so-called quantum computing, which became a very active area of theoretical computer science. Many people hope to construct an effective quantum computer soon.

**3.4.1. Unitary spaces and mappings.** The definitions of scalar product and orthogonality easily extend to the complex case. But we do not mean the complex bilinear symmetric forms  $\alpha$ , since there the quadratic expressions  $\alpha(v, v)$  are not real in general and thus we would not get the right definition of length of vectors. Instead, we define:



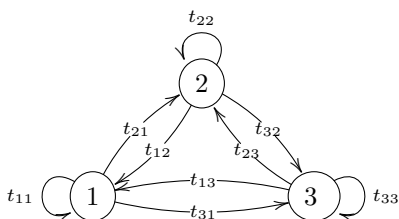
#### UNITARY SPACES

*Unitary space* is a complex vector space  $V$  along with the mapping  $V \times V \rightarrow \mathbb{C}$ ,  $(u, v) \mapsto u \cdot v$  called *scalar product* and satisfying for all vectors  $u, v, w \in V$  and scalars  $a \in \mathbb{C}$  the following axioms:

- (1)  $u \cdot v = \overline{v \cdot u}$  (the bar stands for complex conjugation),
- (2)  $(au) \cdot v = a(u \cdot v)$ ,
- (3)  $(u + v) \cdot w = u \cdot w + v \cdot w$ ,
- (4) if  $u \neq 0$ , then  $u \cdot u > 0$  (notice  $u \cdot u$  is always real).

The real number  $\sqrt{v \cdot v}$  is called the *norm of the vector*  $v$  and a vector is *normalized*, if its norm equals one. Vectors  $u$  and  $v$  are said to be *orthogonal* if their scalar product is zero. A basis composed of mutually orthogonal and normalized vectors is called an *orthonormal basis* of  $V$ .

As for the transition graph it has the following form:



For the final task compute the eigenvector of  $T$  corresponding to the dominant eigenvalue 1. It is given by  $(y_1 = 2, y_2 = 1, y_3 = 1)^T$ , and we should rescale it to obtain the desired probability. This equals to  $y_3/(y_1 + y_2 + y_3) = 1/4$ .  $\square$



Let  $\{X_n : n \geq 0\}$  be a (homogeneous) Markov process with transition matrix  $T = (t_{ij})$  and let  $x^{(n)}$  be the probability distribution of the states at the  $n$ th transition, i.e., the vector whose  $i$ th component describes the probability of the system to be in state  $i$  after  $n$  steps:  $x_i^{(n)} = P(X_n = i)$ . We have  $\sum_i x_i^{(n)} = 1$  and it follows that each  $x^{(n)}$  is a stochastic vector. Moreover, in terms of the matrix  $T$  we know that  $x^{(n+1)} = Tx^{(n)}$ , which gives  $x^{(n)} = T^n x^{(0)}$ . This makes the Markov process computationally very tractable, since once  $T$  is known one can use matrix calculus to compute  $T^n$ . Let us describe such applications.

**3.C.13. Experiment in a laboratory.** In a laboratory an experiment is carried on with the same probability of success and failure. If the experiment succeeds, the probability of success of the second experiment is 0.7. If the first experiment fails, the probability of success of the second experiment is 0.6. This process is continued indefinitely. For any  $n \in \mathbb{N}$  determine the probability that the  $n$ th experiment is successful.

**Solution.** This is a two-state Markov process  $\{X_n : n \in \mathbb{N}\}$  with transition matrix

$$T = \begin{pmatrix} 7/10 & 3/5 \\ 3/10 & 2/5 \end{pmatrix}.$$

The corresponding state space admits the description {success, failure} which here we denote by  $\{1, 2\}$ . Now, for  $n \in \mathbb{N}$  consider the stochastic vectors  $x^{(n)} = (x_1^{(n)}, x_2^{(n)})^T$ , where  $x_1^{(n)}$  is the probability of the success of the  $n$ th experiment and  $x_2^{(n)}$  is the probability of its failure, that is,

$$\begin{aligned} x_1^{(n)} &= P(X_n = 1), \\ x_2^{(n)} &= P(X_n = 2) = 1 - x_1^{(n)}. \end{aligned}$$

By assumption we have  $x^{(1)} = (1/2, 1/2)^T$ , and by the theoretical result in 3.3.5 we obtain the relation

$$x^{(2)} = Tx^{(1)} = (13/20, 7/20)^T.$$

Similarly, for general  $n \in \mathbb{N}$  and the relation  $x^{(n+1)} = Tx^{(n)}$ , we obtain  $x^{(n+1)} = T^2 x^{(n-1)} = \dots = T^n x^{(1)}$ .

At first sight this is an extension of the definition of Euclidean vector spaces into the complex domain. We will continue to use the alternative notation  $\langle u, v \rangle$  for the scalar product of vectors  $u$  and  $v$ . As in the real domain, we obtain immediately from the definition the following simple properties of the scalar product for all vectors in  $V$  and scalars in  $\mathbb{C}$ :

$$\begin{aligned} u \cdot u &\in \mathbb{R} \\ u \cdot u &= 0 \quad \text{if and only if} \quad u = 0 \\ u \cdot (av) &= \bar{a}(u \cdot v) \\ u \cdot (v + w) &= u \cdot v + u \cdot w \\ u \cdot 0 &= 0 \cdot u = 0 \\ \left(\sum_i a_i u_i\right) \cdot \left(\sum_j b_j v_j\right) &= \sum_{i,j} a_i \bar{b}_j (u_i \cdot v_j), \end{aligned}$$

where the last equality holds for all finite linear combinations. It is a simple exercise to prove everything formally. For instance, the first property follows from (1) since the product  $u \cdot u$  has to be the complex conjugate to itself.

A standard example of the scalar product over the complex vector space  $\mathbb{C}^n$  is

$$(x_1, \dots, x_n)^T \cdot (y_1, \dots, y_n)^T = x_1 \bar{y}_1 + \dots + x_n \bar{y}_n.$$

This expression is also called the standard (positive definite) Hermitian form on  $\mathbb{C}^n$ . By conjugation of the coordinates of the second argument, this mapping satisfies all the required properties. The space  $\mathbb{C}^n$  with this scalar product is called the *standard unitary space* of dimension  $n$ . We can denote this scalar product of vectors  $x$  and  $y$  with matrix notation as  $\bar{y}^T \cdot x$  (here the complex conjugation indicated by the bar is performed on all components of  $y$ ).

As usual, those mappings which leave the additional structure invariant are of great importance.

#### UNITARY MAPPINGS

A linear mapping  $\varphi : V \rightarrow W$  between unitary spaces is called a *unitary mapping*, if for all vectors  $u, v \in V$

$$u \cdot v = \varphi(u) \cdot \varphi(v).$$

*Unitary isomorphism* is a bijective unitary mapping.



**3.4.2. Real and complex spaces with scalar product.** In the previous chapter we have already derived some simple properties of spaces with scalar products. The properties and proofs are very similar to the complex case.

In the sequel we shall work with real and complex spaces simultaneously and write  $\mathbb{K}$  for  $\mathbb{R}$  or  $\mathbb{C}$ . In the real case the conjugation is just the identity mapping (it is the restriction of the conjugation in the complex plane to the real line). As in the real case, we define the *orthogonal complement* for a vector subspace  $U \subset V$  in the unitary space  $V$  as

$$U^\perp = \{v \in V; u \cdot v = 0 \text{ for all } u \in U\},$$

which is clearly also a vector subspace in  $V$ .

Next we compute  $T^n$ . By running the following command in Sage we verify that  $T$  is diagonalizable and obtain its eigenvalues: The block

```
A=matrix(QQ, [[7/10, 3/5], [3/10, 2/5]])
A.eigenvalues()
```

Thus, the built-in function `A.is_diagonalizable()` provides a straightforward method to determine if a matrix  $A$  is diagonalizable. We also obtain the eigenvalues  $\lambda_1 = 1$  and  $\lambda_2 = 0.1$ .

As eigenvectors we choose the vectors  $e_1 = (2, 1)$  and  $e_2 = (-1, 1)$ . Since  $T$  is diagonalizable, we can compute  $T^n$  by applying the same method as in 3.B.6. Hence we should compute  $T^n = PD^nP^{-1}$ , where  $P = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}$  and  $D = \text{diag}(\lambda_1, \lambda_2) = \text{diag}(1, 1/10)$ . This gives

$$T^n = \frac{1}{3} \begin{pmatrix} 2 + 10^{-n} & 2 - 2 \cdot 10^{-n} \\ 1 - 10^{-n} & 1 + 2 \cdot 10^{-n} \end{pmatrix}, \quad n \in \mathbb{N}.$$

Here is a verification in Sage:

```
T=matrix(SR, [[7/10, 3/5], [3/10, 2/5]])
P=matrix(SR, [[2, -1], [1, 1]])
n=var("n")
D=matrix(SR, [[1**n, 0], [0, (1/10)**n]])
Tn=P*D*P.inverse(); Tn
```

Now, matrix multiplication of  $T^n$  with the vector  $x^{(1)}$  gives

$$x^{(n+1)} = \left( \frac{2}{3} - \frac{1}{6 \cdot 10^n}, \frac{1}{3} + \frac{1}{6 \cdot 10^n} \right)^T, \quad n \in \mathbb{N}.$$

Thus, for big  $n$  the probability of success of the  $n$ th experiment is close to  $2/3$ . In other words, for large  $n$  we have  $x_1^{(n+1)} \approx 2/3$ .  $\square$

To summarize, another significant advantage of studying Markov chains is their ability to predict the long-term behaviour of a system. Let us illustrate this fact with a straightforward example, related to forecasting weather pattern.



**3.C.14. Weather expectations.** Suppose that the days are divided into warm, medium, and cold, and that the following hold:

1. After a warm day the next day is warm with possibility 50%, and is medium with possibility 30%;
2. After a medium day the next day is medium with possibility 40%, and cold with possibility 30%;
3. After a cold day, the next day is cold with possibility 50% and medium with possibility 30%.

Without any further information, derive how many warm, medium and cold days can be expected in a year.

**Solution.** The problem is clearly a Markov process, since it is assumed that the daily weather depends only on the weather of the previous day.

Although we deal exclusively with finitely-dimensional spaces now, the results in the next two theorems have a natural generalization for Hilbert spaces, which are infinitely-dimensional spaces with scalar products. We shall meet them later, in connection with approximation in vector spaces of real or complex valued functions.

**Theorem.** For every finitely-dimensional space  $V$  of dimension  $n$  with scalar product we have:

- (1) There exists an orthonormal basis in  $V$ .
- (2) Every system of non-zero orthogonal vectors in  $V$  is linearly independent and can be extended to an orthogonal basis.
- (3) For every system of linearly independent vectors  $(u_1, \dots, u_k)$  there exists an orthonormal basis  $(v_1, \dots, v_n)$  such that  $\langle v_1, \dots, v_i \rangle = \langle u_1, \dots, u_i \rangle$ , for all  $1 \leq i \leq k$ , i.e. its vectors consecutively generate the same subspaces as the vector  $u_j$ .
- (4) If  $(u_1, \dots, u_n)$  is an orthonormal basis  $V$ , then the coordinates of every vector  $u \in V$  are expressed via

$$u = (u \cdot u_1)u_1 + \dots + (u \cdot u_n)u_n.$$

- (5) In any orthonormal basis, the scalar product has the coordinate form

$$u \cdot v = \bar{y}^T \cdot x = x_1\bar{y}_1 + \dots + x_n\bar{y}_n$$

where  $x$  and  $y$  are columns of coordinates of the vectors  $u$  and  $v$  in a chosen basis. Notably, every  $n$ -dimensional space with scalar product is isomorphic to the standard Euclidean  $\mathbb{R}^n$  or the unitary  $\mathbb{C}^n$ .

- (6) The orthogonal sum of unitary subspaces  $V_1 + \dots + V_k$  in  $V$  is always a direct sum.
- (7) If  $A \subset V$  is an arbitrary subset, then  $A^\perp \subset V$  is a vector subspace (and thus also unitary), and  $(A^\perp)^\perp \subset V$  is exactly the subspace generated by  $A$ . Furthermore,  $V = \text{span } A \oplus A^\perp$ .
- (8)  $V$  is an orthogonal sum of  $n$  one-dimensional unitary subspaces.

**PROOF.** (1), (2), (3): First we extend the given system of vectors into any basis  $(u_1, \dots, u_n)$  of the space  $V$  and then start the Gram-Schmidt orthogonalization from 2.3.21. This procedure works in the complex case. It yields an orthogonal basis with properties as required in (3). But from the Gram-Schmidt orthogonalization algorithm it is clear that if the original  $k$  vectors formed an orthogonal system of vectors, then they continue to do so after the orthogonalization process is applied. Thus we have also proved (2) and (1).

- (4): If  $u = a_1u_1 + \dots + a_nu_n$ , then

$$u \cdot u_i = a_1(u_1 \cdot u_i) + \dots + a_n(u_n \cdot u_i) = a_i \|u_i\|^2 = a_i$$

- (5): If  $u = x_1u_1 + \dots + x_nu_n, v = y_1u_1 + \dots + y_nu_n$ , then

$$\begin{aligned} u \cdot v &= (x_1u_1 + \dots + x_nu_n) \cdot (y_1u_1 + \dots + y_nu_n) \\ &= x_1\bar{y}_1 + \dots + x_n\bar{y}_n. \end{aligned}$$



This has a 3-dimensional state space  $S = \{w, m, c\}$ , where  $w$  is for warm,  $m$  for medium and  $c$  for cold., and state transition matrix given by  $T = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.3 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$ .

Let us now consider the probabilistic vector  $x^{(n)} = (x_w^{(n)}, x_m^{(n)}, x_c^{(n)})^T$ , whose components are the probabilities the  $n$ th day to be warm, medium or cold. We indicate them by the numbers  $x_w^{(n)} = P(X_n = w)$ ,  $x_m^{(n)} = P(X_n = m)$ , and  $x_c^{(n)} = P(X_n = c)$ , respectively. Since all the elements  $T$  are positive, there exists a probabilistic vector  $x_\infty = (x_\infty^w, x_\infty^m, x_\infty^c)^T$ , which the vector  $x^{(n)}$  approaches as  $n$  grows. By the corollary of the Perron-Frobenius theorem (see 3.3.4),  $x_\infty$  must be the eigenvector of  $T$  corresponding to the eigenvalue 1. This gives rise to the condition  $Tx_\infty = x_\infty$ , which together with the condition that the vector  $x_\infty$  is stochastic, yield the following system of equations:

$$\begin{aligned} x_\infty^w &= 0.5 x_\infty^w + 0.3 x_\infty^m + 0.2 x_\infty^c, \\ x_\infty^m &= 0.3 x_\infty^w + 0.4 x_\infty^m + 0.3 x_\infty^c, \\ x_\infty^c &= 0.2 x_\infty^w + 0.3 x_\infty^m + 0.5 x_\infty^c, \\ 1 &= x_\infty^w + x_\infty^m + x_\infty^c. \end{aligned}$$

It is easy to see that  $x_\infty^w = x_\infty^m = x_\infty^c = \frac{1}{3}$  is the unique solution of this system. Thus, one should expect roughly the same number of warm, medium and cold days.  $\square$

The column-stochastic matrix  $T$  in Problem 3.C.14 is symmetric,  $T^T = T$ . Consequently, it is also “row-stochastic”, indicating that the sum of the entries in any row is one. A matrix which is both column- and row-stochastic is called a “doubly stochastic matrix”. A significant property of every doubly stochastic primitive matrix is that the corresponding vector  $x_\infty$  has all its components identical, as above. This means that, after sufficiently many iterations, all states in the corresponding Markov chain are reached with the same frequency.

A series of additional problems related to Markov chains is presented in Section F. In this section we address the development of an algorithm for determining the importance of web pages (see 3.F.20), along with other compelling tasks.

### D. More matrix calculus

In this section we explore fundamental concepts of linear algebra that underpin advanced applications across diverse fields. We begin with the notion of “unitary spaces”, see also 3.4.1.

Given the vector space  $V = \mathbb{F}^n$ , where  $\mathbb{F}$  is either  $\mathbb{R}$  or  $\mathbb{C}$ , consider the standard scalar product  $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{F}$  defined by  $\langle x, y \rangle := \sum_j x_j \bar{y}_j = x^T \bar{y}$ , where  $x = (x_1, \dots, x_n)^T$  and  $y = (y_1, \dots, y_n)^T$  are vectors in  $\mathbb{F}^n$ , and the last expression is in terms of matrices. Since  $x^T \bar{y} \in \mathbb{F}$  is a scalar, we have  $x^T \bar{y} = (x^T \bar{y})^T = \bar{y}^T x = y^* x$ ,



(6): We need to show that for any tuple  $V_i, V_j$  from the given subspaces their intersection is the zero vector. If  $u \in V_i$  and  $u \in V_j$ , then  $u \perp u$ , that is,  $u \cdot u = 0$ . This is possible only for the zero vector  $u \in V$ .

(7): Let  $u, v \in A^\perp$ . Then  $(au + bv) \cdot w = 0$  for all  $w \in A$ ,  $a, b \in \mathbb{K}$  (from the distributivity of the scalar product). Thus  $A^\perp$  is a subspace in  $V$ . Let  $(v_1, \dots, v_k)$  be a basis of  $\text{span } A$  chosen among the elements of  $A$ , and let  $(u_1, \dots, u_k)$  be the orthonormal basis resulting from the Gram-Schmidt orthogonalization of the vectors  $(v_1, \dots, v_k)$ . We extend it to an orthonormal basis of the whole  $V$  (both exist by the already proven parts of this proposition). Because it is an orthogonal basis, necessarily  $\text{span}\{u_{k+1}, \dots, u_n\} = \text{span}\{u_1, \dots, u_k\}^\perp = A^\perp$  and  $A \subset \text{span}\{u_{k+1}, \dots, u_n\}^\perp$  (this follows from expressing the coordinates under the orthonormal basis). If  $u \perp \text{span}\{u_{k+1}, \dots, u_n\}$ , then  $u$  is necessarily a linear combination of the vectors  $u_1, \dots, u_k$ , but that happens whenever it is a linear combination of the vectors  $v_1, \dots, v_k$ , which is equivalent to  $u$  being in  $\text{span } A$ .

(8): This is equivalent to the formulation of the existence of the orthonormal basis.  $\square$

### 3.4.3. Important properties of the norm.

Now we have everything prepared for basic properties related to our definition of the norm of vectors. We speak also of the *length of vectors* defined by the scalar product. Note also that all claims always consider finite sets of vectors, Their validity does not depend on the dimension of the space  $V$  where it all takes place.



#### PROPERTIES OF NORM

**Theorem.** Let  $V$  be a vector space with scalar product,  $u$  and  $v$  vectors in  $V$ . Then

(1)  $\|u + v\| \leq \|u\| + \|v\|$ . Equality holds if and only if  $u$  and  $v$  are linearly dependent. This is called the **triangle inequality**.

(2)  $|u \cdot v| \leq \|u\| \|v\|$ . Equality holds if and only if  $u$  and  $v$  are linearly dependent. This property is called the **Cauchy inequality**.

(3) If  $(e_1, \dots, e_k)$  is an orthonormal system of vectors, then  $\|u\|^2 \geq |u \cdot e_1|^2 + \dots + |u \cdot e_k|^2$ .

This property is called the **Bessel inequality**.

(4) If  $(e_1, \dots, e_k)$  is an orthonormal system of vectors, then  $u \in \text{span}\{e_1, \dots, e_k\}$  if and only if

$$\|u\|^2 = |u \cdot e_1|^2 + \dots + |u \cdot e_k|^2.$$

This is called the **Parseval equality**.

(5) If  $(e_1, \dots, e_k)$  is an orthonormal system of vectors and  $u \in V$ , then the vector

$$w = (u \cdot e_1)e_1 + \dots + (u \cdot e_k)e_k$$

is the only vector which minimizes the norm  $\|u - v\|$  among all  $v \in \text{span}\{e_1, \dots, e_k\}$ .

where  $y^* := \bar{y}^T$ . Hence, we can equivalently express  $\langle \cdot, \cdot \rangle$  as

$$\langle x, y \rangle = x^T \bar{y} = y^* x.$$

By convention,  $\langle \cdot, \cdot \rangle$  is linear in the first argument but conjugate linear in the second one, in the sense that  $\langle ax, y \rangle = a \langle x, y \rangle$  but  $\langle x, ay \rangle = \bar{a} \langle x, y \rangle$ , for all  $x, y \in V$  and  $a \in \mathbb{F}$ .

The scalar product  $\langle \cdot, \cdot \rangle$  induces a norm on  $\mathbb{F}^n$  defined by  $\|x\|^2 = \langle x, x \rangle = \sum_{j=1}^n |x_j|^2$ .<sup>11</sup> Ensure that  $\|\cdot\|$  satisfies the defining properties of a norm, that is,

- $\|x\| \geq 0$  with  $\|x\| = 0$  if and only if  $x = 0$ ,
- $\|ax\| = |a| \|x\|$ , for any  $a \in \mathbb{F}$  and  $x \in \mathbb{F}^n$
- $\|x + y\| \leq \|x\| + \|y\|$ , for any  $x, y \in \mathbb{F}^n$ .

Next we will adopt the following terminology:

$$\langle \cdot, \cdot \rangle = \begin{cases} \text{standard dot product,} & \text{for } V = \mathbb{R}^n, \\ \text{standard Hermitian form,} & \text{for } V = \mathbb{C}^n, \end{cases}$$

This notation is consistent with what is used in Chapter 2 for the real case. Both spaces  $\mathbb{R}^n, \mathbb{C}^n$  endowed with  $\langle \cdot, \cdot \rangle$  provide examples of unitary spaces (also known as inner product spaces), as defined in 3.4.1. On the other hand, the distance map  $d : V \times V \rightarrow \mathbb{R}$ , defined by  $d(x, y) = \|x - y\|$ , establishes  $V = \mathbb{F}^n$  as a “metric space”, a fundamental concept in analysis which we will explore in Chapter 7

Next, we will see that unitary spaces extend beyond  $\mathbb{R}^n$  and  $\mathbb{C}^n$ , with many other examples existing. We begin with the following task, which is left as an easy challenge for you.

**3.D.1.** Consider the vector space  $\mathbb{C}^3$ , endowed with the standard Hermitian form  $\langle x, y \rangle = \sum_k x_k \bar{y}_k$  and the induced norm  $\|\cdot\|$ . Given the vectors  $x = (3 + 2i, 1 - i, -i)^T$  and  $y = (2 - 2i, 1 - i, 2 + i)^T$ , compute the inner product  $\langle x, y \rangle$ , the distance  $d(x, y) = \|x - y\|$ , and the normalized vectors  $\hat{x}, \hat{y}$  corresponding to  $x, y$ , with  $\|\hat{x}\| = 1 = \|\hat{y}\|$ .  $\circ$

**3.D.2.** When using Sage to compute scalar products of complex vectors, we need to be cautious. Sage permits the use of the standard command `x.dot_product(y)`, as discussed in Chapter 2 (see for example 2.C.45), to complex vectors as well. However, the handling of complex vectors may require special attention to ensure correct application and interpretation of the standard Hermitian scalar product.

For example, let us use the vectors  $x, y$  given in 3.D.1. Execute the cell

```
x = vector(CDF, [3+2*I, 1-I, -I])
y = vector(CDF, [2-2*I, 1-I, 2+I])
y.dot_product(x)
```

In this case Sage prints out the expression  $11.0 - 6.0 * I$ , which does not match the result obtained using the standard rule  $\langle x, y \rangle = \sum_i x_i \bar{y}_i$ . To obtain the correct result we should use the cell

<sup>11</sup>If  $z_j = x_j + iy_j \in \mathbb{C}$  recall that  $\bar{z}_j = x_j - iy_j$  and  $|z_j|^2 = x_j^2 + y_j^2$ .

**PROOF.** The verifications are all based on direct computations:

(2): The result is obvious if  $v = 0$ . Otherwise, define the vector  $w = u - \frac{u \cdot v}{v \cdot v} v$ , that is,  $w \perp v$  and compute

$$\begin{aligned} \|w\|^2 &= \|u\|^2 - \frac{(u \cdot v)^2}{\|v\|^2} - \frac{u \cdot v}{\|v\|^2} (v \cdot u) + \frac{(u \cdot v)(\overline{u \cdot v})}{\|v\|^4} \|v\|^2 \\ \|w\|^2 \|v\|^2 &= \|u\|^2 \|v\|^2 - 2(u \cdot v)(\overline{u \cdot v}) + (u \cdot v)(\overline{u \cdot v}) \end{aligned}$$

These are non-negative real values and thus,  $\|u\|^2 \|v\|^2 \geq |u \cdot v|^2$  and the equality holds if and only if  $w = 0$ , that is, whenever  $u$  and  $v$  are linearly dependent.

(1): It suffices to compute

$$\begin{aligned} \|u + v\|^2 &= \|u\|^2 + \|v\|^2 + u \cdot v + v \cdot u \\ &= \|u\|^2 + \|v\|^2 + 2 \operatorname{Re}(u \cdot v) \\ &\leq \|u\|^2 + \|v\|^2 + 2|u \cdot v| \leq \|u\|^2 + \|v\|^2 + 2\|u\| \|v\| \\ &= (\|u\| + \|v\|)^2 \end{aligned}$$

Since we deal with squares of non-negative real numbers, this means that  $\|u + v\| \leq \|u\| + \|v\|$ . Furthermore, equality implies that in all previous inequalities equality also holds. This is equivalent to the condition that  $u$  and  $v$  are linearly dependent (using the previous part).

(3), (4): Let  $(e_1, \dots, e_k)$  be an orthonormal system of vectors. We extend it to an orthonormal basis  $(e_1, \dots, e_n)$  (that is always possible by the previous theorem). Then, again using the previous theorem, we have for every vector  $u \in V$

$$\|u\|^2 = \sum_{i=1}^n (u \cdot e_i)(\overline{u \cdot e_i}) = \sum_{i=1}^n |u \cdot e_i|^2 \geq \sum_{i=1}^k |u \cdot e_i|^2$$

But that is the Bessel inequality. Furthermore, equality holds if and only if  $u \cdot e_i = 0$  for all  $i > k$ , which proves the Parseval equality.

(5): Choose an arbitrary  $v \in \operatorname{span}\{e_1, \dots, e_k\}$  and extend the given orthonormal system to the orthonormal basis  $(e_1, \dots, e_n)$ . Let  $(u_1, \dots, u_n)$  and  $(x_1, \dots, x_k, 0, \dots, 0)$  be coordinates of  $u$  and  $v$  under this basis. Then

$$\|u - v\|^2 = |u_1 - x_1|^2 + \dots + |u_k - x_k|^2 + |u_{k+1}|^2 + \dots + |u_n|^2$$

and this expression is clearly minimized when choosing the individual vectors to be  $x_1 = u_1, \dots, x_k = u_k$ .  $\square$

**3.4.4. Unitary and orthogonal mappings.** The properties of orthogonal mappings have direct analogues in the complex domain. We can easily formulate them and prove together:



**Proposition.** Consider the linear mapping (endomorphism)  $\varphi : V \rightarrow V$  on the (real or complex) space with scalar product. Then the following conditions are equivalent.

- (1)  $\varphi$  is unitary or orthogonal transformation,
- (2)  $\varphi$  is linear isomorphism and for every  $u, v \in V$

$$\varphi(u) \cdot v = u \cdot \varphi^{-1}(v),$$

- (3) the matrix  $A$  of the mapping  $\varphi$  in any orthonormal basis satisfies  $A^{-1} = \bar{A}^T$  (for Euclidean spaces this means that  $A^{-1} = A^T$ ),

```
x = vector(CDF, [3+2*I, 1-I, -I])
y = vector(CDF, [2-2*I, 1-I, 2+I])
y.hermitian_inner_product(x)
```

which returns  $3.0 + 8.0 * I$ .

**Remark.** It is important to note that in Sage the command `x.hermitian_inner_product(y)` prints out the expression  $3.0 - 8.0 * I$ . Therefore, Sage uses the rule  $\sum_i \bar{x}_i y_i$ , instead of our rule  $\sum_i x_i \bar{y}_i$ . In summary, when computing the standard Hermitian form of complex vectors  $x, y \in \mathbb{C}^n$ , to stay consistent with our conventions, we should use the rule

$$\langle x, y \rangle = y^* x = y.\text{hermitian\_inner\_product}(x).$$

An alternative is based on the `dot.product` function and goes as follows:

```
u.dot_product(v.conjugate())
```

This corresponds to  $\langle u, v \rangle$  for two complex vectors  $u, v$ .

**3.D.3.** Use the standard Hermitian form  $\langle v, w \rangle = \sum_i v_i \bar{w}_i$  to compute  $\langle v, w \rangle$ ,  $\|v\|^2$ ,  $\|w\|^2$  and the angle  $\theta$  between  $v, w$ , where these vectors are given as follows:

(a)  $v = (1 + i, 2 - i)^T, w = (3 - 2i, 1 + i)^T$  in  $\mathbb{C}^2$ ;

(b)  $v = (3, i)^T, w = (2 - i, 1 - i)^T$  in  $\mathbb{C}^2$ ;

(c)  $v = (-i, 0, 2)^T, w = (4, 1 - i, 1)^T$  in  $\mathbb{C}^3$ ;

Next verify your answers via Sage. ○

**3.D.4.** On  $V = \mathbb{F}^2$  for  $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$  consider the map

$$f((x_1, x_2)^T, (y_1, y_2)^T) = x_1 \bar{y}_1 + 4x_1 \bar{y}_2 + 4x_2 \bar{y}_1 + x_2 \bar{y}_2.$$

Does  $f$  define a scalar product on  $V$ ? ○

**3.D.5.** For  $x = (x_1, x_2)^T$  and  $y = (y_1, y_2)^T$  in  $\mathbb{R}^2$  set

$$g(x, y) := 2x_1 y_1 - x_1 y_2 - x_2 y_1 + 5x_2 y_2.$$

Show that  $g$  is a scalar product and compute its matrix relative to the standard basis of  $\mathbb{R}^2$ .

**Solution.** Linearity and symmetry can be easily proved. In addition, we see that

$$g(x, x) = 2x_1^2 - 2x_1 x_2 + 5x_2^2 = (x_1 + x_2)^2 + (x_1 - 2x_2)^2,$$

for all  $x \in \mathbb{R}$ . Thus  $g(x, x) \geq 0$ , in particular  $g(x, x) = 0$  if and only if  $x_1 + x_2 = 0$  and  $x_1 - 2x_2 = 0$ , which gives  $x = 0$ . Thus  $g$  is a scalar product. It is clearly different from the standard dot product. Its matrix with respect to the standard basis of  $\mathbb{R}^2$  is given by

$$A = \begin{pmatrix} g(e_1, e_1) & g(e_1, e_2) \\ g(e_2, e_1) & g(e_2, e_2) \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 5 \end{pmatrix},$$

such that  $g(x, y) = y^T A x$ . Observe, that  $A$  is symmetric and positive definite (recall that an element  $A \in \text{Mat}_n(\mathbb{R})$  is called a “positive definite matrix” when we have  $u^T A u > 0$  for any non-zero vector  $u \neq 0$  on  $\mathbb{R}^n$ ). □

**3.D.6.** Let  $a \neq b$  be some positive real numbers. Show that the rule  $\rho_{a,b}(u, v) := au_1 v_1 + bu_2 v_2$  defines a scalar product on  $\mathbb{R}^2$ , where  $u = (u_1, u_2)^T$  and  $v = (v_1, v_2)^T$ , respectively. Next, compute the angle between the vectors  $u = (1, 1)^T$  and

(4) the matrix  $A$  of a mapping  $\varphi$  in some orthonormal basis satisfies  $A^{-1} = \bar{A}^T$ ,

(5) the rows of the matrix  $A$  of the mapping  $\varphi$  in an orthonormal basis form an orthonormal basis of the space  $\mathbb{K}^n$  with standard scalar product,

(6) the columns of the matrix  $A$  of the mapping  $\varphi$  in an orthonormal basis form an orthonormal basis of the space  $\mathbb{K}^n$  with standard scalar product.

**PROOF.** (1)  $\Rightarrow$  (2): The mapping  $\varphi$  is injective, therefore it must be onto. Also  $\varphi(u) \cdot v = \varphi(u) \cdot \varphi(\varphi^{-1}(v)) = u \cdot \varphi^{-1}(v)$ .



(2)  $\Rightarrow$  (3): The standard scalar product is in  $\mathbb{K}^n$ . It is given for columns  $x, y$  of scalars by the expression  $x \cdot y = \bar{y}^T E x = \bar{y} x$ , where  $E$  is the unit matrix.

Property (2) thus means that the matrix  $A$  of the mapping  $\varphi$  is invertible and  $\bar{y}^T A x = (\bar{A}^{-1} y)^T x$ . This means that  $(\bar{y}^T A - (\bar{A}^{-1} y)^T) x = 0$  for all  $x \in \mathbb{K}^n$ . By substituting the complex conjugate of the expression in the parentheses for  $x$  we find that equality is possible only when  $\bar{A}^T = A^{-1}$ . (We may also rewrite the expression as  $\bar{y}^T (A - (\bar{A}^{-1})^T) x$  and see the conclusion by substituting the basis vectors for  $x$  and  $y$ .)

(3)  $\Rightarrow$  (4): This is an obvious implication.

(4)  $\Rightarrow$  (5) In the relevant basis, the claim is expressed via the matrix  $A$  of the mapping  $\varphi$  as the equation  $A \bar{A}^T = E$ , which is ensured by (4).

(5)  $\Rightarrow$  (6): We have  $|\bar{A}^T A| = |E| = |A \bar{A}^T| = |A| |\bar{A}| = 1$ , there exists the inverse matrix  $A^{-1}$ . But we also have  $A \bar{A}^T A = A$ , therefore also  $\bar{A}^T A = E$  which is expressed exactly by (6).

(6)  $\Rightarrow$  (1): In the chosen orthonormal basis

$$\varphi(u) \cdot \varphi(v) = (\overline{Ay})^T Ax = \bar{y}^T \bar{A}^T Ax = \bar{y}^T E x = \bar{y}^T x$$

where  $x$  and  $y$  are columns of coordinates of the vectors  $u$  and  $v$ . That ensures that the scalar product is preserved. □

Characterizations from the previous theorem deserve some notes. The matrices  $A \in \text{Mat}_n(\mathbb{K})$  with the property  $A^{-1} = \bar{A}^T$  are called *unitary matrices* for complex scalars (in the case  $\mathbb{R}$  we have already used the name *orthogonal matrices* for them). The definition itself immediately implies that a product of unitary (orthogonal) matrices is again unitary (orthogonal). The same is true for inverses. Unitary matrices thus form a subgroup  $U(n) \subset \text{GL}_n(\mathbb{C})$  in the group of all invertible complex matrices with the product operation. Orthogonal matrices form a subgroup  $O(n) \subset \text{GL}_n(\mathbb{R})$  in the group of real invertible matrices. We speak of a *unitary group* and of an *orthogonal group*.


The simple calculation

$$1 = \det E = \det(A \bar{A}^T) = \det A \overline{\det A} = |\det A|^2$$

shows that the determinant of a unitary matrix has norm equal to one. For real scalars the determinant is  $\pm 1$ . Furthermore, if  $Ax = \lambda x$  for a unitary or orthogonal matrix, then

$v = (1, -1)^T$  with respect to  $\rho_{2,1}$  and compare the result with the angle that occurs if we use the dot product on  $\mathbb{R}^2$ .  $\circ$

**3.D.7.** Suppose that  $A = (a_{ij})$  is a  $m \times n$  matrix over  $\mathbb{C}$ , whose column space has (complex) dimension  $n$ . Consider the mapping  $\rho_A : \mathbb{C}^n \times \mathbb{C}^n \rightarrow \mathbb{C}$  with  $\rho_A(x, y) = y^* A^* A x$ , where as usual  $A^* = \bar{A}^T$  denotes the transpose conjugate of a matrix  $A$ . Show that the pair  $(\mathbb{C}^n, \rho_A)$  is a unitary space.  $\circ$

**3.D.8.** Show that the rule  $\mathcal{B}(A, B) := \text{tr}(B^* A)$  defines a scalar product on the space  $\text{Mat}_{m,n}(\mathbb{F})$  of  $m \times n$  matrices with entries over  $\mathbb{F}$ , for  $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ . This scalar product is known as the *Frobenius inner product*. 

**Solution.** For any  $A, B \in \text{Mat}_{m,n}(\mathbb{F})$ , the matrix  $B^* A$  is  $n \times n$ , and hence  $\mathcal{B}$  is well-defined. One can proceed by proving the axioms of a unitary space, as they are presented in 3.4.1. However, a direct calculation yields that

$$\mathcal{B}(A, B) = \sum_{i=1}^m \sum_{j=1}^n \bar{b}_{ij} a_{ij} = \sum_{i=1}^m \sum_{j=1}^n a_{ij} \bar{b}_{ij}$$

Hence, if we express  $A$  and  $B$  in terms of column vectors, say  $A = (A_1 \ \dots \ A_n)$  and  $B = (B_1 \ \dots \ B_n)$ , then  $\mathcal{B}(A, B) = \sum_{j=1}^n \langle A_j, B_j \rangle$ , which is the sum of the dot products of the vectors formed from the columns of the matrices  $A, B$ , respectively. Hence  $\mathcal{B}$  is a scalar product (as a sum of dot products). Otherwise, a direct proof is mainly based on the properties of the trace, described in 2.C.38 in Chapter 2. For instance, over the complex numbers we have  $\mathcal{B}(aA, B) = \text{tr}(B^*(aA)) = a \text{tr}(B^* A) = a\mathcal{B}(A, B)$  for any scalar  $a \in \mathbb{C}$  and any two elements  $A, B \in \text{Mat}_{m,n}(\mathbb{C})$ , and moreover, if  $C \in \text{Mat}_{m,n}(\mathbb{C})$  is another matrix, then

$$\begin{aligned} \mathcal{B}(A + C, B) &= \text{tr}(B^*(A + C)) = \text{tr}(B^* A) + \text{tr}(B^* C) \\ &= \mathcal{B}(A, B) + \mathcal{B}(C, B). \end{aligned}$$

Positive-definiteness occurs as follows: If  $A^* A$  has entries  $c_{ij}$ , then  $c_{ij} = \sum_{k=1}^n \bar{a}_{ki} a_{kj}$  and thus

$$\|A\|_{\mathcal{B}}^2 = \mathcal{B}(A, A) = \sum_{i=1}^n c_{ii} = \sum_{i,j} |a_{ij}|^2.$$

Hence, as soon as  $A \neq 0$ , we see that  $\mathcal{B}(A, A)$  is strictly positive,  $\mathcal{B}(A, A) > 0$ . We leave as an exercise a direct verification of the property  $\mathcal{B}(A, B) = \bar{\mathcal{B}}(B, A)$  for  $A, B \in \text{Mat}_{m,n}(\mathbb{C})$ .  $\square$

*In Chapter 7 we will encounter additional examples of unitary spaces. There, leveraging the concept of integration, which we will examine in Chapter 6, we will introduce inner products on spaces of polynomials and, more generally, on infinite-dimensional function spaces (see for example 7.1.1, 7.1.2 and see also the tasks 7.D.3, 7.D.4, 7.D.5, and 7.D.6). In this context, orthogonality becomes crucial, especially in Fourier analysis. For now, let us explore a few more elementary tasks related to orthogonality.*

$(Ax) \cdot (Ax) = x \cdot x = |\lambda|^2(x \cdot x)$ . Therefore the real eigenvalues of orthogonal matrices in the real domain are  $\pm 1$ . The eigenvalues of unitary matrices are always complex units in the complex plane.

The same argument as we have seen with the orthogonal mappings imply that orthogonal complements of invariant subspaces with respect to unitary mappings  $\varphi : V \rightarrow V$  are also invariant. Indeed, if  $\varphi(U) \subset U$ ,  $u \in U$  and  $v \in U^\perp$  are arbitrary, then

$$\varphi(v) \cdot \varphi(\varphi^{-1}(u)) = v \cdot \varphi^{-1}(u).$$


Because the restriction  $\varphi|_U$  is also unitary, it is a bijection. Notably  $\varphi^{-1}(u) \in U$ . But then  $\varphi(v) \cdot u = 0$ , because  $v \in U^\perp$ . Thus  $\varphi(v) \in U^\perp$ .

This leads to an immediate useful corollary in the complex domain

**Corollary.** *Let  $\varphi : V \rightarrow V$  be a unitary mapping of complex vector spaces. Then  $V$  is an orthogonal sum of one-dimensional eigensubspaces.*

**PROOF.** There exists at least one eigenvector  $v \in V$ , since complex eigenvalues always exist. Then the restriction of  $\varphi$  to the invariant subspace  $\langle v \rangle^\perp$  is again unitary and also has an eigenvector. After  $n$  such steps we obtain the desired orthogonal basis of eigenvectors. After normalising the vectors we obtain an orthonormal basis.  $\square$

Now it is possible to understand the details of the proof of the spectral decomposition of the orthogonal mapping from 2.4.7 at the end of the second chapter. The real matrix of an orthogonal mapping is interpreted as a matrix of a unitary mapping on a complex extension of Euclidean space. We observe the corollaries of the structure of the roots of the real characteristic polynomial over the complex domain. Automatically we obtain invariant two-dimensional subspaces given by pairs of complex conjugated eigenvalues and hence the corresponding rotation for restricted original real mapping.

**3.4.5. Dual and adjoint mappings.** When discussing vector spaces and linear mappings in the second chapter, we mentioned briefly the dual vector space  $V^*$  of all linear forms over the vector space  $V$ , see 2.3.17. This duality extends to mappings: 

DUAL MAPPINGS

For any linear mapping  $\psi : V \rightarrow W$ , the expression

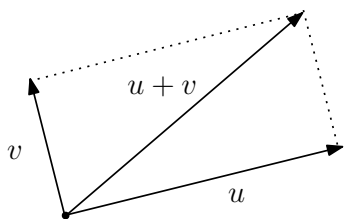
$$(1) \quad \langle v, \psi^*(\alpha) \rangle = \langle \psi(v), \alpha \rangle,$$

where  $\langle \cdot, \cdot \rangle$  denotes the evaluation of the linear forms (the second argument) on the vectors (the first argument), while  $v \in V$  and  $\alpha \in W^*$  are arbitrary, defines the mapping  $\psi^* : W^* \rightarrow V^*$  called the *dual mapping* to  $\psi$ .

Choose bases  $\underline{v}$  in  $V$ ,  $\underline{w}$  in  $W$  and write  $A$  for the matrix of the mapping  $\psi$  in these bases. Then we compute the matrix of the mapping  $\psi^*$  in the corresponding dual bases in the



**3.D.9. (The Pythagorean theorem)** (a) In a real vector space  $V$  with a scalar product  $\langle \cdot, \cdot \rangle$  prove that two vectors  $u, w$  are orthogonal, i.e.,  $\langle u, w \rangle = 0$ , if and only if  $\|u+w\|^2 = \|u\|^2 + \|w\|^2$ .<sup>12</sup>



Next demonstrate with a counterexample that this property does not hold for a complex unitary vector space.

(b) If  $(V, \langle \cdot, \cdot \rangle)$  is a real scalar vector space, prove that the vectors  $u - w$  and  $u + w$  are orthogonal, if and only if  $\|u\| = \|w\|$ , where  $u, w \in V$  are two arbitrary vectors.  $\circ$

**3.D.10.** Let  $(V, \langle \cdot, \cdot \rangle)$  be a unitary space over  $\mathbb{F}$ , and  $u, v \in V$  two arbitrary vectors. Show that

(a)  $u \perp v$  if and only if  $\|u + av\| = \|u - av\|$  for all  $a \in \mathbb{F}$ .

(b)  $u \perp v$  if and only if  $\|u + av\| \geq \|u\|$  for all  $a \in \mathbb{F}$ .  $\circ$

**3.D.11.** Consider the space  $\text{Mat}_{m,n}(\mathbb{R})$  with the scalar product  $\mathcal{B}(A, B) = \text{tr}(B^T A)$ , introduced in 3.D.8. For the matrices

$$A = \begin{pmatrix} 1 & 3 & 5 \\ 0 & 2 & 2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 & 4 & 0 \\ 7 & 9 & 1 \end{pmatrix}$$

compute the following:

(a) the angle  $\theta$  between  $A, B$ ;

(b) the distance  $\|A - B\|_{\mathcal{B}}$  between  $A, B$ ;

(c) Verify the Cauchy-Schwarz inequality.

**Solution.** (a) By 3.D.8, for two real matrices  $A = (a_{ij})$  and  $B = (b_{ij})$ , both of size  $m \times n$ , we get  $\mathcal{B}(A, B) = \text{tr}(B^T A) = \sum_{i=1}^m \sum_{j=1}^n a_{ij} b_{ij}$  and  $\|A\|_{\mathcal{B}}^2 = \mathcal{B}(A, A) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2$ . In particular, if  $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$

and  $B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix}$  are two elements of  $\text{Mat}_{2,3}(\mathbb{R})$ ,

then we see that the product  $B^T A$  is given by

$$\begin{pmatrix} b_{11}a_{11} + b_{21}a_{21} & b_{11}a_{12} + b_{21}a_{22} & b_{11}a_{13} + b_{21}a_{23} \\ b_{12}a_{11} + b_{22}a_{21} & b_{12}a_{12} + b_{22}a_{22} & b_{12}a_{13} + b_{22}a_{23} \\ b_{13}a_{11} + b_{23}a_{21} & b_{13}a_{12} + b_{23}a_{22} & b_{13}a_{13} + b_{23}a_{23} \end{pmatrix}$$

such that

$$\begin{aligned} \mathcal{B}(A, B) &= \text{tr}(B^T A) = a_{11}b_{11} + a_{12}b_{12} + a_{13}b_{13} \\ &\quad + a_{21}b_{21} + a_{22}b_{22} + a_{23}b_{23}. \end{aligned}$$

Thus for the given  $A$  and  $B$  we compute

$$\begin{aligned} \mathcal{B}(A, B) &= 2 + 12 + 0 + 0 + 18 + 2 = 34, \\ \|A\|_{\mathcal{B}}^2 &= a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{21}^2 + a_{22}^2 + a_{23}^2 = 43, \\ \|B\|_{\mathcal{B}}^2 &= b_{11}^2 + b_{12}^2 + b_{13}^2 + b_{21}^2 + b_{22}^2 + b_{23}^2 = 151. \end{aligned}$$

<sup>12</sup>Observe that this condition, known as the Pythagorean theorem, is not the same as the equality case in the triangle inequality  $\|u+w\| \leq \|u\| + \|w\|$ .

dual spaces. Indeed, the definition says that if we represent the vectors from  $W^*$  in the coordinates as rows of scalars, then the mapping  $\psi^*$  is given by the same matrix as  $\psi$ , if we multiply by it the row vectors from the right:

$$\langle \psi(v), \alpha \rangle = (\alpha_1, \dots, \alpha_n) \cdot A \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \langle v, \psi^*(\alpha) \rangle.$$

This means that the matrix of the dual mapping  $\psi^*$  is the transpose  $A^T$ , because  $\alpha \cdot A = (A^T \cdot \alpha^T)^T$ .

Assume further that we have a vector space with scalar product. Then we can naturally identify  $V$  and  $V^*$  using the scalar product. Indeed, choosing one fixed vector  $w \in V$ , we substitute this vector into the second argument in the scalar product in order to obtain the identification  $V \simeq V^* = \text{Hom}(V, \mathbb{K})$

$$V \ni w \mapsto (v \mapsto \langle v, w \rangle) \in V^*.$$

The non-degeneracy condition on the scalar product ensures that this mapping is a bijection. Notice it is important to use  $w$  as the fixed second argument in the case  $\mathbb{K} = \mathbb{C}$  in order to obtain linear forms. Since factorizing complex multiples in the second argument yields complex conjugated scalars, the identification  $V \simeq V^*$  is linear over real scalars only.

It is clear that the vectors of an orthonormal basis are mapped to forms that constitute the dual basis, i.e. the orthonormal basis are selfdual under our identification. Moreover, every vector is automatically understood as a linear form, by means of the scalar product.

How does the above dual mapping  $W^* \rightarrow V^*$  look in terms of our identification? We use the same notation  $\psi^* : W \rightarrow V$  for the resulting mapping, which is uniquely given as follows:

ADJOINT MAPPING

For every linear mapping  $\psi : V \rightarrow W$  between spaces with scalar products, there is the *adjoint mapping*  $\psi^*$  uniquely determined by the formula

$$(2) \quad \langle \psi(u), v \rangle = \langle u, \psi^*(v) \rangle.$$

The parentheses means the scalar products on  $W$  or  $V$ , respectively.

Notice that the use of the same parenthesis for evaluation of one-forms and scalar products (which reflects the identification above) makes the defining formulae of dual and adjoint mappings look the same.

Equivalently we can understand the relation (2) to be the definition of the adjoint mapping  $\psi^*$ . By substituting all pairs of vectors from an orthonormal basis for the vectors  $u$  and  $v$  we obtain directly all the values of the matrix of the mapping  $\psi^*$ .



Hence  $\|A\|_{\mathcal{B}} = \sqrt{43}$ ,  $\|B\|_{\mathcal{B}} = \sqrt{151}$ , and

$$\cos \theta = \frac{\mathcal{B}(A, B)}{\|A\|_{\mathcal{B}}\|B\|_{\mathcal{B}}} = \frac{34}{\sqrt{43}\sqrt{151}}.$$

From this one can explicitly compute  $\theta$ , as before.

(b) We see that  $A - B = \begin{pmatrix} -1 & -1 & 5 \\ -7 & -7 & 1 \end{pmatrix}$ . Thus

$$\|A - B\|_{\mathcal{B}}^2 = \mathcal{B}(A - B, A - B) = 1 + 1 + 25 + 49 + 49 + 1 = 126$$

such that  $\|A - B\|_{\mathcal{B}} = \sqrt{126}$ .

(c) For the Cauchy inequality we refer to 3.4.3. This important inequality for the scalar product  $\mathcal{B}$  takes the form

$$|\mathcal{B}(A, B)| \leq \|A\|_{\mathcal{B}}\|B\|_{\mathcal{B}} = \sqrt{\mathcal{B}(A, A)}\sqrt{\mathcal{B}(B, B)} \iff |\text{tr}(B^* A)| \leq \sqrt{\text{tr}(A^* A)}\sqrt{\text{tr}(B^* B)}.$$

This gives us  $34 < \sqrt{43}\sqrt{151} \approx 80$ .  $\square$



We proceed with exercises related to orthonormal bases and orthogonal complements. Recall that we have already discussed these concepts in Chapter 2 (see for example 2.C.48, 2.C.50). There we introduced the Gram-Schmidt orthogonalization process, which transforms any basis  $\{E_1, \dots, E_n\}$  of a scalar product space  $(V, \langle \cdot, \cdot \rangle)$ , to an orthogonal basis  $\{w_1, \dots, w_n\}$  of  $V$ .

The process begins with  $w_1 = E_1$ , and constructs the  $j$ th member  $w_j$  as follows:

$$w_j = E_j - \frac{\langle E_j, w_1 \rangle}{\|w_1\|^2} w_1 - \dots - \frac{\langle E_j, w_{j-1} \rangle}{\|w_{j-1}\|^2} w_{j-1}.$$

This results to an orthonormal basis  $\{\frac{w_1}{\|w_1\|}, \dots, \frac{w_n}{\|w_n\|}\}$  of  $V$ .

This straightforward yet essential construction has a wide range of applications. We begin with an example that requires extending the method presented in Chapter 2 (cf. 2.C.50), to fit Sage's capabilities.

**3.D.12.** Consider the Euclidean space  $\mathbb{R}^3$  endowed with the scalar product  $\langle\langle u, v \rangle\rangle = v^T A u$ , where  $A$  is given by<sup>13</sup>

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Applying the Gram-Schmidt procedure to obtain an  $\langle\langle \cdot, \cdot \rangle\rangle$ -orthogonal basis of  $\mathbb{R}^3$ , starting with the standard basis  $e = \{e_1, e_2, e_3\}$  of  $\mathbb{R}^3$ .

**Solution.** Let  $e_1 = (1, 0, 0)^T$ ,  $e_2 = (0, 1, 0)^T$  and  $e_3 = (0, 0, 1)^T$  be the vectors of the standard basis  $e$  of  $\mathbb{R}^3$ . According to the Gram-Schmidt method, an  $\langle\langle \cdot, \cdot \rangle\rangle$ -orthogonal basis of  $\mathbb{R}^3$  is given by  $\{w_1, w_2, w_3\}$ , with  $w_1 = e_1$  and

$$\begin{aligned} w_2 &= e_2 - \frac{\langle\langle e_2, w_1 \rangle\rangle}{\langle\langle w_1, w_1 \rangle\rangle} w_1, \\ w_3 &= e_3 - \frac{\langle\langle e_3, w_1 \rangle\rangle}{\langle\langle w_1, w_1 \rangle\rangle} w_1 - \frac{\langle\langle e_3, w_2 \rangle\rangle}{\langle\langle w_2, w_2 \rangle\rangle} w_2. \end{aligned}$$

<sup>13</sup>Verify yourselves that  $A$  is symmetric and positive definite (for the second claim show for example that all the eigenvalues of  $A$  are positive).

Using the coordinate expression for the scalar product, the formula (2) reveals the coordinate expression of the adjoint mapping:

$$\begin{aligned} \langle\psi(v), w\rangle &= \overline{(w_1, \dots, w_n)} \cdot A \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \\ &= \overline{\left(\bar{A}^T \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}\right)^T} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \langle v, \psi^*(w) \rangle. \end{aligned}$$

It follows that if  $A$  is the matrix of the mapping  $\psi$  in an orthonormal basis, then the matrix of the adjoint mapping  $\psi^*$  is the transposed and conjugated matrix  $A$  – we denote this by  $A^* = \bar{A}^T$ .

The matrix  $A^*$  is called the *adjoint matrix* of the matrix  $A$ . Note that the adjoint matrix is well defined for any rectangular matrix. We should not confuse them with algebraic adjoints, which we used for square matrices when working with determinants.

We can summarise. For any linear mapping  $\psi : V \rightarrow W$  between unitary spaces, with matrix  $A$  in some bases on  $V$  and  $W$ , its dual mapping has the matrix  $A^T$  in the dual basis. If there are scalar products on  $V$  and  $W$ , we identify them (via the scalar products) with their duals. Then the dual mapping coincides with the adjoint mapping  $\psi^* : W \rightarrow V$ , which has the matrix  $A^*$ . The distinction between the matrix of the dual mapping and the matrix of the adjoint mapping is thus in the additional conjugation. This is of course a consequence of the fact that our identification of the unitary space with its dual is not a linear mapping over complex scalars.

**3.4.6. Self-adjoint mappings.** Those linear mappings which coincide with their adjoints:  $\psi^* = \psi$ , are of particular interest. They are called *self-adjoint mappings*. Equivalently we can say that they are the mappings whose matrix  $A$  satisfies  $A = A^*$  in some (and thus in all) orthonormal basis.



In the case of Euclidean spaces the self-adjoint mappings are those with symmetric matrices (in orthonormal basis). They are often called *symmetric mappings*.

In the complex domain the matrices that satisfy  $A = A^*$  are called *Hermitian matrices* or also *Hermitian symmetric matrices*. Sometimes they are also called *self-adjoint matrices*. Note that Hermitian matrices form a real vector subspace in the space of all complex matrices, but it is not a vector subspace in the complex domain.

**Remark.** The next observation is of special interest. If we multiply a Hermitian matrix  $A$  by the imaginary unit, we obtain the matrix  $B = i A$ , which has the property

$$B^* = \bar{i} \bar{A}^T = -B.$$

Such matrices are called *anti-Hermitian* or *Hermitian skew-symmetric*. Every real matrix can be written as a sum of its

We compute

$$\langle\langle e_2, w_1 \rangle\rangle = (1 \ 0 \ 0) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = -1,$$

$$\langle\langle w_1, w_1 \rangle\rangle = (1 \ 0 \ 0) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 2.$$

Thus  $w_2 = e_2 + \frac{1}{2}e_1 = (\frac{1}{2}, 1, 0)^T$ . Next we compute

$$\begin{aligned} \langle\langle e_3, w_1 \rangle\rangle &= \langle\langle e_3, e_1 \rangle\rangle = e_1^T A e_3 \\ &= (1 \ 0 \ 0) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 0, \end{aligned}$$

that is,  $e_1$  and  $e_3$  are  $\langle\langle \cdot, \cdot \rangle\rangle$ -orthogonal, and hence

$$\begin{aligned} \langle\langle e_3, w_2 \rangle\rangle &= \langle\langle e_3, e_2 + \frac{1}{2}e_1 \rangle\rangle = \langle\langle e_3, e_2 \rangle\rangle + \frac{1}{2}\langle\langle e_3, e_1 \rangle\rangle \\ &= \langle\langle e_3, e_2 \rangle\rangle = e_2^T A e_3 \\ &= (0 \ 1 \ 0) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 1. \end{aligned}$$

Moreover,

$$\langle\langle w_2, w_2 \rangle\rangle = (\frac{1}{2} \ 1 \ 0) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ 1 \\ 0 \end{pmatrix} = \frac{3}{2}.$$

As an alternative, we know that  $\langle\langle e_1, e_1 \rangle\rangle = 2$ ,  $\langle\langle e_1, e_2 \rangle\rangle = -1$ , and we compute  $\langle\langle e_2, e_2 \rangle\rangle = 2$ . By linearity, we then have

$$\begin{aligned} \langle\langle w_2, w_2 \rangle\rangle &= \frac{1}{4}\langle\langle e_1, e_1 \rangle\rangle + \langle\langle e_1, e_2 \rangle\rangle + \langle\langle e_2, e_2 \rangle\rangle \\ &= \frac{1}{4} \cdot 2 + (-1) + 2 = \frac{3}{2}. \end{aligned}$$

Hence,  $w_3 = e_3 - \frac{2}{3}w_2 = (-\frac{1}{3}, -\frac{2}{3}, 1) = -\frac{1}{3}e_1 - \frac{2}{3}e_2 + e_3$ . To summarize, the desired basis has the form

$$\left\{ e_1, \frac{1}{2}e_1 + e_2, -\frac{1}{3}e_1 - \frac{2}{3}e_2 + e_3 \right\}.$$

To confirm this basis in Sage we may revise the method presented in Chapter 2 (see 2.C.50). This adjustment is necessary because we are now using a different scalar product than the standard one. The following cell includes the necessary explanations and performs the Gram-Schmidt orthogonalization with respect to the given inner product defined by the matrix  $A$ . Execute this block yourselves to read Sage's result.

symmetric part and its anti-symmetric part,

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T).$$

In the complex domain we have analogously

$$A = \frac{1}{2}(A + A^*) + i\frac{1}{2i}(A - A^*).$$

In particular, we may express every complex matrix in a unique way as a sum

$$A = B + iC$$

with Hermitian symmetric matrices  $B = \frac{1}{2}(A + A^*)$  and  $C = \frac{1}{2i}(A - A^*)$ . This is an analogy of the decomposition of a complex number into its real and purely imaginary component and in the literature we often encounter the notation

$$B = \operatorname{re} A = \frac{1}{2}(A + A^*), \quad C = \operatorname{im} A = \frac{1}{2i}(A - A^*).$$

In the language of linear mappings this means that every complex linear automorphism can be uniquely expressed by means of two self-adjoint mappings playing the role of the real and imaginary parts of the original mapping.

**3.4.7. Spectral decomposition.** Consider a self-adjoint mapping  $\psi : V \rightarrow V$  with the matrix  $A$  in some orthonormal basis. Proceed similarly as in 2.4.7 when we diagonalized the matrix of orthogonal mappings.



Again, consider arbitrary invariant subspaces of self-adjoint mappings and their orthogonal complements. If a self-adjoint mapping  $\psi : V \rightarrow V$  leaves a subspace  $W \subset V$  invariant, i.e.  $\psi(W) \subset W$ , then for every  $v \in W^\perp$ ,  $w \in W$

$$\langle\psi(v), w\rangle = \langle v, \psi(w)\rangle = 0.$$

Thus also,  $\psi(W^\perp) \subset W^\perp$ .

Next, consider the matrix  $A$  of a self-adjoint mapping in an orthonormal basis and an eigenvector  $x \in \mathbb{C}^n$ , i.e.  $A \cdot x = \lambda x$ . We obtain

$$\lambda \langle x, x \rangle = \langle Ax, x \rangle = \langle x, Ax \rangle = \bar{\lambda} \langle x, x \rangle.$$

The positive real number  $\langle x, x \rangle$  can be cancelled on both sides and thus  $\bar{\lambda} = \lambda$ , and we see that eigenvalues of Hermitian matrices are always real.

The characteristic polynomial  $\det(A - \lambda E)$  has as many complex roots as is the dimension of the square matrix  $A$  (including multiplicities), and all of them are actually real. Thus we have proved the important general result:

**Proposition.** *The orthogonal complements of invariant subspaces of self-adjoint mappings are also invariant. Furthermore, the eigenvalues of a Hermitian matrix  $A$  are always real.*

The very definition ensures that restriction of a self-adjoint mapping to an invariant subspace is again self-adjoint. Thus the latter proposition implies that there always exists an orthonormal basis of  $V$  composed of eigenvectors. Indeed, start with any eigenvector  $v_1$ , normalize it, consider its linear hull  $V_1$  and restrict the mapping to  $V_1^\perp$ . Consider next

```
A = Matrix([[2, -1, 0], [-1, 2, 1], [0, 1, 2]])
E1 = vector([1, 0, 0])
E2 = vector([0, 1, 0])
E3 = vector([0, 0, 1])
def proj(v, u):
    p = ((v*A*u) / (u*A*u)) * u
    return p
w1 = E1; w2 = E2 - proj(E2, w1)
w3 = E3 - proj(E3, w2) - proj(E3, w1)
```

another eigenvector  $v_1 \in V_2^\perp$ , take  $V_2 = \text{span}(V_1 \cup \{v_2\})$ , which is again invariant. Continue and construct the sequence of invariant subspaces  $V_1 \subset V_2 \subset \dots \subset V_n = V$ , building the orthonormal basis of eigenvectors, as expected.

Actually, it is easy to see directly that eigenvectors associated with different eigenvalues are perpendicular to each other. Indeed, if  $\psi(u) = \lambda u$ ,  $\psi(v) = \mu v$  then we obtain

$$\lambda \langle u, v \rangle = \langle \psi(u), v \rangle = \langle u, \psi(v) \rangle = \bar{\mu} \langle u, v \rangle = \mu \langle u, v \rangle.$$

Usually this result is formulated using projections onto eigensubspaces. Recall the properties of projections along subspaces, as discussed in 2.3.19. A projection  $P : V \rightarrow V$  is a linear mapping satisfying  $P^2 = P$ . This means that the restriction of  $P$  to its image is the identity and the projector is completely determined by choosing the subspaces  $\text{Im } P$  and  $\text{Ker } P$ .

A projection  $P : V \rightarrow V$  is called *orthogonal* if  $\text{Im } P \perp \text{Ker } P$ . Two orthogonal projections  $P, Q$  are called *mutually perpendicular* if  $\text{Im } P \perp \text{Im } Q$ .

SPECTRAL DECOMPOSITION OF SELF-ADJOINT MAPPINGS

**Theorem** (Spectral decomposition). *For every self-adjoint mapping  $\psi : V \rightarrow V$  on a vector space with scalar product there exists an orthonormal basis composed of eigenvectors. If  $\lambda_1, \dots, \lambda_k$  are all distinct eigenvalues of  $\psi$  and if  $P_1, \dots, P_k$  are the corresponding orthogonal and mutually perpendicular projectors onto the eigenspaces corresponding to the eigenvalues, then*

$$\psi = \lambda_1 P_1 + \dots + \lambda_k P_k.$$

*The dimensions of the images of these projections  $P_i$  equal the algebraic multiplicities of the eigenvalues  $\lambda_i$ .*

**3.4.8. Orthogonal diagonalization.** Linear mappings



which allow for orthonormal bases as in the latter theorem on spectral decomposition are called *orthogonally diagonalizable*. Of course, they are exactly the mappings for which we can find an orthonormal basis in which the matrix of the mapping is diagonal. We ask what they look like.

In the Euclidean case, this is simple: diagonal matrices are first of all symmetric, thus they are the self-adjoint mappings. As a corollary we note that an orthogonal mapping of an Euclidean space into itself is orthogonally diagonalizable if and only if it is self-adjoint. They are exactly the self-adjoint mappings with eigenvalues  $\pm 1$ .

The situation is much more interesting on unitary spaces. Consider any linear mapping  $\varphi : V \rightarrow V$  on a unitary space. Let  $\varphi = \psi + i\eta$  be the (unique) decomposition of  $\varphi$  into its Hermitian and anti-Hermitian part. If  $\varphi$  has diagonal matrix  $D$  in a suitable orthonormal basis, then  $D = \text{Re } D + i \text{Im } D$ , where the real and the imaginary parts are exactly the matrices of  $\psi$  and  $\eta$ . This follows from the uniqueness of the decomposition. Knowing this in the particular coordinates, we conclude the following computation relations at the level of

```
g = [w1, w2, w3]
show(g)
```

□

**3.D.13.** Let  $(V, \langle \cdot, \cdot \rangle)$  be a (finite-dimensional) unitary vector space and let  $\{E_j\}_{j=1}^n$  be an  $\langle \cdot, \cdot \rangle$ -orthonormal basis of  $V$ , i.e.,  $\langle E_i, E_j \rangle = \delta_{ij}$ . Show that any two vectors  $x, y \in V$  satisfy  $\langle x, y \rangle = \sum_{j=1}^n \langle x, E_j \rangle \langle y, E_j \rangle$ . ○

**3.D.14.** Show that the vectors  $E_1 = (-1, 1, 2)$ ,  $E_2 = (2, 0, 1)$  and  $E_3 = (1, 5, -2)$  form an orthogonal basis of  $\mathbb{R}^3$  (with respect to the usual dot product). Next express the vector  $u = (6, 2, -4)$  as a linear combination of this basis. ○

**3.D.15.** Based on the Cauchy-Schwarz inequality, show that any triple  $(a, b, c)$  of positive real numbers satisfies the following inequality:

$$\sqrt{\frac{a+2b}{a+b+c}} + \sqrt{\frac{b+2c}{a+b+c}} + \sqrt{\frac{c+2a}{a+b+c}} \leq 3.$$

○

**3.D.16.** Consider  $\mathbb{C}^3$  endowed with the standard Hermitian form  $\langle u, v \rangle = \sum_{i=1}^3 u_i \bar{v}_i$  and the standard basis  $\underline{e} = \{e_1, e_2, e_3\}$ . Verify Parseval's equality for the vector  $u = (2+i, -1+2i, 3-i) \in \mathbb{C}^3$ . ○

**3.D.17.** In Problem 3.D.11 use the isomorphism  $\text{Mat}_{2,3}(\mathbb{R}) \cong \mathbb{R}^6$  established in Problem 2.C.31 to show that the matrices  $A, B$  are linearly independent. Then consider the subspace  $W$  of  $\text{Mat}_{2,3}(\mathbb{R})$  spanned by the matrices  $A, B$ . Find a basis of the orthogonal complement  $W^\perp$  of  $W$  with respect to the Frobenius scalar product  $\mathcal{B}$  (introduced in 3.D.8).

**Solution.** Under the isomorphism  $\varphi : \text{Mat}_{2,3}(\mathbb{R}) \cong \mathbb{R}^6$  discussed in Problem 2.C.31, we may view the matrices  $A$  and  $B$  as the following vectors on  $\mathbb{R}^6$ :  $\varphi(A) = v_1 = (1, 3, 5, 0, 2, 2)^T$  and  $\varphi(B) = v_2 = (2, 4, 0, 7, 9, 1)^T$ , respectively. We now treat their linear independence via Sage by the cell

```
V = RR^6
v1 = vector(RR, [1, 3, 5, 0, 2, 2])
v2 = vector(RR, [2, 4, 0, 7, 9, 1])
V.linear_dependence([v1, v2]) == []
```

Sage prints out True, so  $v_1, v_2$  are linearly independent. Consider the subspace  $W = \text{span}_{\mathbb{R}}\{A, B\} \cong \text{span}_{\mathbb{R}}\{v_1, v_2\}$  of  $\text{Mat}_{2,3}(\mathbb{R}) \cong \mathbb{R}^6$  spanned by  $A, B$ . We need to determine  $W^\perp$  with respect to  $\mathcal{B}$ , that is,

$$W^\perp = \{C \in \text{Mat}_{2,3}(\mathbb{R}) : \mathcal{B}(C, A) = 0 = \mathcal{B}(C, B)\}.$$

Let us express  $C \in \text{Mat}_{2,3}(\mathbb{R})$  as  $C = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$ , for some reals  $a, \dots, f$ . Then  $C \in W^\perp$  if and only if  $\text{tr}(A^T C) = 0 = \text{tr}(B^T C)$ , that is

$$\{a + 3b + 5c + 2e + 2f = 0, 2a + 4b + 7d + 9e + f = 0\}.$$

mappings  $\psi \circ \eta = \eta \circ \psi$  (i.e. the real and imaginary parts of  $\varphi$  commute), and  $\varphi \circ \varphi^* = \varphi^* \circ \varphi$  (since this clearly holds for all diagonal metrics). The mappings  $\varphi : V \rightarrow V$  with the latter property are called the *normal mappings*.

A detailed characterization is given by the following theorem (stated in the notation of this paragraph):

**Theorem.** The following conditions on a mapping  $\varphi : V \rightarrow V$  on a unitary space  $V$  are equivalent:

- (1)  $\varphi$  is orthogonally diagonalizable,
- (2)  $\varphi^* \circ \varphi = \varphi \circ \varphi^*$  ( $\varphi$  is a normal mapping),
- (3)  $\psi \circ \eta = \eta \circ \psi$  (the Hermitian and anti-Hermitian parts commute),
- (4) if  $A = (a_{ij})$  is the matrix of  $\varphi$  in some orthonormal basis, and  $\lambda_i$  are the  $m = \dim V$  eigenvalues of  $A$ , then

$$\sum_{i,j=1}^m |a_{ij}|^2 = \sum_{i=1}^m |\lambda_i|^2.$$

**PROOF.** The implication (1)  $\Rightarrow$  (2) was discussed above.

(2)  $\Leftrightarrow$  (3): it suffices to calculate

$$\varphi \circ \varphi^* = (\psi + i\eta)(\psi - i\eta) = \psi^2 + \eta^2 + i(\eta\psi - \psi\eta)$$

$$\varphi^* \circ \varphi = (\psi - i\eta)(\psi + i\eta) = \psi^2 + \eta^2 + i(\psi\eta - \eta\psi)$$

Subtraction of the two lines yields

$$\varphi\varphi^* - \varphi^*\varphi = 2i(\eta\psi - \psi\eta).$$

(2)  $\Rightarrow$  (1): If  $\varphi$  is normal, then

$$\begin{aligned} \langle \varphi(u), \varphi(u) \rangle &= \langle \varphi^* \varphi(u), u \rangle = \langle \varphi \varphi^*(u), u \rangle \\ &= \langle \varphi^*(u), \varphi^*(u) \rangle \end{aligned}$$

thus  $|\varphi(u)| = |\varphi^*(u)|$ .

Next, notice  $(\varphi - \lambda \text{id } V)^* = (\varphi^* - \bar{\lambda} \text{id } V)$ . Thus, if  $\varphi$  is normal, then  $(\varphi - \lambda \text{id } V)$  is normal too.

If  $\varphi(u) = \lambda u$ , then  $u$  is in the kernel of  $\varphi - \lambda \text{id } V$ . Thus the latter equality of norms of values for normal mappings and their adjoints ensures that  $u$  is also in the kernel of  $\varphi^* - \bar{\lambda} \text{id } V$ . It follows that  $\varphi^*(u) = \bar{\lambda}u$ . We have proved, under the assumption (2), that  $\varphi$  and  $\varphi^*$  have the same eigenvectors and that they are associated to conjugated eigenvalues.

Similarly to our procedure with self-adjoint mappings, we now prove orthogonal diagonalizability. The latter procedure is based on the fact that the orthogonal complements to sums of eigenspaces are invariant subspaces.

Consider an eigenvector  $u \in V$  with eigenvalue  $\lambda$ , and any  $v \in \langle u \rangle^\perp$ . We have

$$\langle \varphi(v), u \rangle = \langle v, \varphi^*(u) \rangle = \langle v, \bar{\lambda}u \rangle = \lambda \langle u, v \rangle = 0.$$

Thus  $\varphi(v) \in \langle u \rangle^\perp$ . The same occurs if  $u$  is replaced by a sum of eigenvectors instead.

(1)  $\Rightarrow$  (4): the expression  $\sum_{i,j} |a_{ij}|^2$  is the trace of the matrix  $AA^*$ , which is the matrix of the mapping  $\varphi \circ \varphi^*$ . Therefore its value does not depend on the choice of the orthonormal basis. Thus if  $\varphi$  is diagonalizable, this expression equals exactly  $\sum_i |\lambda_i|^2$ .

(4)  $\Rightarrow$  (1): This part of the proof is a direct corollary of the Schur theorem on unitary triangulation of an arbitrary

This is a system of two equations with six unknowns. Let  $c, d, e, f \in \mathbb{R}$  be the free variables. Then we get the solution

$$a = 10c - \frac{21}{2}d - \frac{19}{2}e + \frac{5}{2}f, \quad b = -5c + \frac{7}{2}d + \frac{5}{2}e - \frac{3}{2}f.$$

Hence  $W^\perp$  consists of matrices of the form

$$C = \begin{pmatrix} 10c - \frac{21}{2}d - \frac{19}{2}e + \frac{5}{2}f & -5c + \frac{7}{2}d + \frac{5}{2}e - \frac{3}{2}f & c \\ & d & e \\ & & f \end{pmatrix}$$

with  $c, d, e, f \in \mathbb{R}$ . We see that

$$\begin{aligned} C &= c \begin{pmatrix} 10 & -5 & 1 \\ 0 & 0 & 0 \end{pmatrix} + d \begin{pmatrix} -21/2 & 7/2 & 0 \\ 1 & 0 & 0 \end{pmatrix} \\ &\quad + e \begin{pmatrix} -19/2 & 5/2 & 0 \\ 0 & 1 & 0 \end{pmatrix} + f \begin{pmatrix} 5/2 & -3/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= cW_1 + dW_2 + eW_3 + fW_4, \end{aligned}$$

where we denote the matrices appearing above by  $W_1, W_2, W_3, W_4$ , respectively. This shows that  $W_1, \dots, W_4$  generate  $W^\perp$ . Actually, they provide a basis of  $W^\perp$ . For a quick verification of their linear independence, we use again the isomorphism  $\text{Mat}_{2,3}(\mathbb{R}) \cong \mathbb{R}^6$  and proceed with Sage, as before:

```
V = RR^6
w1 = vector(RR, [10, -5, 1, 0, 0, 0])
w2 = vector(RR, [-21/2, 7/2, 0, 1, 0, 0])
w3 = vector(RR, [-19/2, 5/2, 0, 0, 1, 0])
w4 = vector(RR, [5/2, -3/2, 0, 0, 0, 1])
V.linear_dependence([w1, w2, w3, w4]) == []
```

□

*Invertible linear transformations naturally intersect with group theory, a topic that we briefly introduced in Chapter 1 and will explore in greater detail in Chapter 12. This intersection forms the basis of “matrix groups”, which are of particular interest in this context. Next we will focus on the group of all invertible linear mappings from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ , known as the “real general linear group”, denoted by  $\text{GL}_n(\mathbb{R})$ .*

*The operation that defines this group is the composition of linear mappings. Equivalently,  $\text{GL}_n(\mathbb{R})$  can be described as the group of all invertible  $n \times n$  matrices with real entries, where the group operation is the matrix multiplication, i.e.,*

$$\text{GL}_n(\mathbb{R}) = \{A \in \text{Mat}_n(\mathbb{R}) : \det(A) \neq 0\}.$$

*On the other hand, a “matrix group” is defined as a closed subgroup of  $\text{GL}_n(\mathbb{R})$ .<sup>14</sup> An example of a matrix group*

<sup>14</sup>Matrix groups are special cases of the well-known “Lie groups”. In simple terms, a Lie group is a group equipped with a compatible differentiable structure, also known as a smooth manifold. Lie groups were introduced by the Norwegian mathematician Sophus Lie (1842-1899) during the late 19th century, shortly after the discovery of non-Euclidean geometries. Lie referred to them as “continuous symmetry groups”. During this period, Lie collaborated with F. Klein, and together they significantly altered perspectives in geometry and the theory of differential equations. Today, the theory of Lie groups and Lie algebras has become a fundamental area in differential geometry with a wide range of applications. It’s worth noting that not all Lie groups are matrix groups, indicating the potential complexity of these structures.

linear mapping  $V \rightarrow V$ , which we prove later in 3.4.15. This theorem says that for every linear mapping  $\varphi : V \rightarrow V$  there exists an orthonormal basis under which  $\varphi$  has an upper triangular matrix. Then all the eigenvalues of  $\varphi$  appear on its diagonal. Since we have already shown that the expression  $\sum_{i,j} |a_{ij}|^2$  does not depend on the choice of the orthonormal bases, all elements in the upper triangular matrix, which are not on the diagonal must be zero. □

**Remark.** We can rephrase the main statement of the latter theorem in terms of matrices. A mapping is normal if and only if its matrix  $A$  satisfies  $AA^* = A^*A$  in some orthonormal basis (and equivalently in any orthonormal basis). Such matrices are called *normal*. Moreover, we can consider the last theorem as a generalization of standard calculations with complex numbers. The linear mappings appear similar to complex numbers in their algebraic form. The role of real numbers is played by self-adjoint mappings, and the unitary mappings play the role of the complex units  $\cos t + i \sin t \in \mathbb{C}$ . The following consequence of the theorem shows the link to the property  $\cos^2 t + \sin^2 t = 1$ .

**Corollary.** *The unitary mappings on a unitary space  $V$  are exactly those normal mappings  $\varphi$  on  $V$  for which the unique decomposition  $\varphi = \psi + i\eta$  into Hermitian and anti-Hermitian parts satisfies  $\psi^2 + \eta^2 = \text{id}_V$ .*

**PROOF.** If  $\varphi$  is unitary, then  $\varphi\varphi^* = \text{id}_V = \varphi^*\varphi$  and thus  $\varphi\varphi^* = (\psi + i\eta)(\psi - i\eta) = \psi^2 + 0 + \eta^2 = \text{id}_V$ . On the other hand, if  $\varphi$  is normal, we can read the latter computation backwards which proves the other implication. □

**3.4.9. Roots of matrices.** Non-negative real numbers are exactly those which are squares of real numbers (and thus we may find their square roots). At the same time, their positive square roots are uniquely defined. Now we observe a similar behaviour of matrices of the form  $B = A^*A$ . Of course, these are the matrices of the compositions of mappings  $\varphi$  with their adjoints.

By definition,

$$(1) \quad \langle Bx, x \rangle = \langle A^*Ax, x \rangle = \langle Ax, Ax \rangle \geq 0$$

for all vectors  $x$ . Furthermore, we clearly have

$$B^* = (A^*A)^* = A^*A = B.$$

Hermitian matrices  $B$  with the property  $\langle Bx, x \rangle \geq 0$  for all  $x$  are called *positive semidefinite* matrices. If the zero value is attained only for  $x = 0$ , they are called *positive definite*. Analogously, we speak of *positive definite* and *positive semidefinite* (self-adjoint) mappings  $\varphi : V \rightarrow V$ .

For every mapping  $\varphi : V \rightarrow V$  we can define its *square root* as a mapping  $\psi$  such that  $\psi \circ \psi = \varphi$ . The next theorem completely describes the situation when restricting to positive semidefinite mappings.



is the “orthogonal group”  $O(n)$ , which consists of all linear transformations  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  of  $\mathbb{R}^n$  preserving the standard Euclidean product, i.e.,  $\langle \varphi(u), \varphi(v) \rangle = \langle u, v \rangle$ , for all  $u, v \in \mathbb{R}^n$ . We studied such endomorphisms in the end of Chapter 2 and we learned that they correspond to orthogonal matrices (see 2.4.6 and 2.D.11). Hence, in terms of matrices,  $O(n)$  consists of all  $n \times n$  orthogonal matrices, i.e.,  $O(n) = \{A \in \text{Gl}_n(\mathbb{R}) : A^{-1} = A^T\}$ .

**3.D.18.** Prove that  $O(n)$  is a group and a subgroup of  $\text{Gl}_n(\mathbb{R})$ . Additionally, demonstrate that the determinant of any matrix  $A \in O(n)$  is either 1 or  $-1$ .

**Solution.** Obviously, the  $n \times n$  identity matrix  $E$  belongs to  $O(n)$  and this is the corresponding identity element of the group. Thus, to verify that  $O(n)$  is a group it remains to prove that the composition of two orthogonal transformations is orthogonal, and that the inverse of an orthogonal transformation is again orthogonal. By the conclusion in Problem 2.D.11, one can equivalently work with orthogonal matrices. Let  $A, B \in O(n)$ . Then we see that

$$\begin{aligned} (AB)^T AB &= B^T A^T AB = B^T B = E, \\ AB(AB)^T &= ABB^T A^T = AA^T = E, \\ (A^{-1})^T A^{-1} &= (A^T)^{-1} A^{-1} = (AA^T)^{-1} = E^{-1} = E, \\ A^{-1}(A^{-1})^T &= A^{-1}(A^T)^{-1} = (A^T A)^{-1} = E^{-1} = E, \end{aligned}$$

and these relations certify our claim. Finally recall that the matrix multiplication is associative.

Now, given a group  $(G, \circ)$ , a non-empty subset  $K$  of  $G$  which is closed under composition and taking inverses (with respect to the restriction of the group operation  $\circ$  to  $K$ ), is called a *subgroup* of  $G$ , see also 12.4.1 for more details. To demonstrate that a (non-empty) subset  $K \subset G$  of  $G$  is a subgroup of  $G$ , we need to show that  $a \circ b^{-1} \in K$ , for any two elements  $a, b \in K$ . Let  $A, B \in O(n)$  be two orthogonal matrices. By the previous assertion we know that  $B^{-1} \in O(n)$  as well, and hence  $AB^{-1} \in O(n)$ . Since  $O(n)$  is also a subset of  $\text{Gl}_n(\mathbb{R})$ , our claim follows. For the determinant, let  $A \in \text{Mat}_n(\mathbb{R})$  be an orthogonal matrix, i.e.,  $AA^T = E = A^T A$ . Hence  $\det(AA^T) = \det(E) = 1$ . However,  $\det(AA^T) = \det(A) \det(A^T) = \det(A)^2$ , thus  $\det(A)^2 = 1$ , i.e.,  $\det(A) = \pm 1$ .  $\square$

The special orthogonal group  $\text{SO}(n)$  is the subgroup of  $O(n)$  consisting of orthogonal matrices with determinant one,  $\text{SO}(n) = \{A \in O(n) : \det(A) = 1\}$ . Obviously, for  $n = 1$  we have  $O(1) \cong \mathbb{Z}_2$  and the group  $\text{SO}(1)$  is trivial. For  $n = 2$  it is not hard to prove that the group  $\text{SO}(2)$  is isomorphic (as a group) to the unit circle  $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ . For  $n = 3$  we have already described examples of special orthogonal matrices in Chapter 2, e.g., rotations on  $\mathbb{R}^3$  about an axis through the origin, which indeed have matrices lying in the special orthogonal group  $\text{SO}(3)$  (see 2.D.13). Let us further emphasize on rotations on the 3-dimensional Euclidean

POSITIVE SEMIDEFINITE SQUARE ROOTS

**Theorem.** For each positive semidefinite square matrix  $B$ , there is the uniquely defined positive semidefinite square root  $\sqrt{B}$ .

If  $P$  is any matrix such that  $P^{-1}BP = D$  is diagonal, then  $\sqrt{B} = P\sqrt{D}P^{-1}$ , where  $D$  has got the (non-negative) eigenvalues of  $B$  on its diagonal and  $\sqrt{D}$  is the matrix with the positive square roots of these values on its diagonal.

**PROOF.** Since  $B$  is a matrix of a self-adjoint mapping  $\varphi$ , there is even an orthonormal  $P$  as in the theorem (cf. Theorem 3.4.7) with all eigenvalues in the diagonal of  $D$  non-negative. Consider  $C = \sqrt{B}$  as defined in the second claim and notice that indeed

$$C^2 = P\sqrt{D}P^{-1}P\sqrt{D}P^{-1} = PDP^{-1} = B.$$

Thus the mapping  $\psi$  given by  $C$  must have the same eigenvectors as  $\varphi$  and thus these two mappings share the decompositions of  $\mathbb{K}^n$  into mutually orthogonal eigenspaces. In particular, both of them will share the bases in which they have diagonal matrices and thus the definition of  $\sqrt{D}$  must be unique in each such basis. This proves that the definition of  $\sqrt{B}$  does not depend on our particular choice of the diagonalization of  $\varphi$ .  $\square$

Notice there could be a lot of different roots, if we relax the positivity condition on  $\sqrt{B}$  (e.g., we may choose the signs in the diagonal matrix  $D$ ).

**3.4.10. Spectra and nilpotent mappings.** We return to the behavior of linear mappings in full generality. We continue to work with real or complex vector spaces, but without necessarily fixing a scalar product there.

Recall that the *spectrum of a linear mapping*  $f : V \rightarrow V$  is a sequence of roots of the characteristic polynomial of the mapping  $f$ , counting multiplicities. The *algebraic multiplicity* of an eigenvalue is its multiplicity as a root of the characteristic polynomial. The *geometric multiplicity* of an eigenvalue is the dimension of the corresponding subspace of eigenvectors.

A linear mapping  $f : V \rightarrow V$  is called *nilpotent*, if there exists an integer  $k \geq 1$  such that the iterated mapping  $f^k$  is identically zero. The smallest  $k$  with such a property is called the *degree of nilpotency* of the mapping  $f$ . The mapping  $f : V \rightarrow V$  is called *cyclic*, if there exists the basis  $(u_1, \dots, u_n)$  of the space  $V$  such that  $f(u_1) = 0$  and  $f(u_i) = u_{i-1}$  for all  $i = 2, \dots, n$ . In other words, the matrix of  $f$  in this basis is of the form

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

space, by presenting some addition exercises, see also in Section F for further material (cf. 3.F.35, 3.F.37).

**3.D.19.** Write down the matrices of rotations by the angle  $\theta$ , about the (oriented) axes  $x$ ,  $y$  and  $z$  in  $\mathbb{R}^3$ .<sup>15</sup> ○



The concept of linear transformations on real vector spaces naturally extends to complex vector spaces. In particular, the groups  $O(n)$  and  $SO(n)$  have counterparts in the complex case, known as the “unitary group”  $U(n)$  and the “special unitary group”  $SU(n)$ . These groups are closed subgroups of the complex general linear group  $GL_n(\mathbb{C})$ , which serves as the starting point in this context, replacing  $GL_n(\mathbb{R})$ . As a result, they are also matrix groups (note that when identifying  $\mathbb{C}^n \cong \mathbb{R}^{2n}$ , the group  $GL_n(\mathbb{C})$  can be viewed as a subgroup of  $GL_{2n}(\mathbb{R})$ ).

The unitary group consists of all complex linear mappings  $\psi : \mathbb{C}^n \rightarrow \mathbb{C}^n$ , preserving the standard Hermitian form  $\langle \cdot, \cdot \rangle$  on  $\mathbb{C}^n$ . Equivalently, it can be viewed as the matrix group  $U(n) = \{U \in GL_n(\mathbb{C}) : U^{-1} = U^* = \bar{U}^T\}$ . Just as  $SO(n)$  is defined as a subgroup of  $O(n)$ , the special unitary group  $SU(n)$  is defined as the subgroup of  $U(n)$  consisting of unitary matrices with determinant one.

**3.D.20.** As we said above, over  $\mathbb{C}$  a  $n \times n$  unitary matrix  $U$  is defined as the one preserving the standard Hermitian form  $\langle \cdot, \cdot \rangle$  on  $\mathbb{C}^n$ , i.e.,  $\langle Ux, Uy \rangle = \langle x, y \rangle$  for any  $x, y \in \mathbb{C}^n$ . Show that this is equivalent to say that  $U^*U = UU^* = E$ . Next prove that  $U(n)$  is a group, and that the determinant of any unitary matrix is a complex unit. ○

**3.D.21. Unitary matrices.** Determine which of the matrices listed below is unitary:

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \end{pmatrix}, B = \frac{1}{2} \begin{pmatrix} 1+i & \sqrt{2} \\ 1-i & \sqrt{2}i \end{pmatrix}, C = \begin{pmatrix} 1 & -i \\ -1 & i \end{pmatrix}.$$

**Solution.** Let us use Sage to analyze the matrix  $A$ . We want to determine whether the products  $A^*A$  and  $AA^*$  yield the identity matrix. We this goal in mind, we type:

```
A=matrix(SR, [[(1/sqrt(2)), (1/sqrt(2))],
[(1/sqrt(2))*I, (-1/sqrt(2))*I]])
show(A)
A_her=A.conjugate_transpose()
show(A_her)
bool(A_her*A==A*A_her)
```

The command `A.conjugate_transpose()` returns the conjugate transpose of  $A$ , and has as shortcut the rule `A.H`. Executing this block we confirm that  $A$  is a unitary matrix. As mentioned before, a more direct method uses the command

<sup>15</sup>The matrices presented in the solution are well-known and are commonly used in 3D graphics, robotics, and other fields involving three-dimensional transformations.

If  $f(v) = av$ , then  $f^k(v) = a^k \cdot v$  for every natural  $k$ . Note that, the spectrum of nilpotent mapping can contain only the zero scalar (and this is always present).

By the definition, every cyclic mapping is nilpotent. Moreover, its degree of nilpotency equals the dimension of the space  $V$ . The derivative operator on polynomials,  $D(x^k) = kx^{k-1}$ , is an example of a cyclic mapping on the spaces  $\mathbb{K}_n[x]$  of all polynomials of degree at most  $n$  over the scalars  $\mathbb{K}$ .

Perhaps surprisingly, this is also true the other way round – every nilpotent mapping is a direct sum of cyclic mappings. A proof of this claim takes much work. So we formulate first the results we are aiming at, and only then come back to the technical work.

In the resulting theorem describing the *Jordan decomposition*, the crucial role is played by vector (sub)spaces and linear mappings with a single eigenvalue  $\lambda$  given by the matrix

$$(1) \quad J = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & \lambda \end{pmatrix}.$$

These matrices (and the corresponding invariant subspaces) are called *Jordan blocks*.<sup>5</sup>

JORDAN CANONICAL FORM

**Theorem.** Let  $V$  be a real or complex vector space of dimension  $n$ . Let  $f : V \rightarrow V$  be a linear mapping with  $n$  eigenvalues (in the chosen domain of scalars), counting algebraic multiplicities. Then there exists a unique decomposition of the space  $V$  into the direct sum of subspaces

$$V = V_1 \oplus \dots \oplus V_k$$

where not only  $f(V_i) \subset V_i$ , but the restriction of  $f$  to each  $V_i$  has a single eigenvalue  $\lambda_i$  and the restriction  $f - \lambda_i \text{id}_{V_i}$  on  $V_i$  is either cyclic or is the zero mapping. In particular, there is a suitable basis in which  $f$  has a block-diagonal matrix  $J$  with Jordan blocks along the diagonal.

We say that the matrix  $J$  from the theorem is in Jordan canonical form. In the language of matrices, we can rephrase the theorem as follows:

**Corollary.** For each square matrix  $A$  over complex scalars, there is an invertible matrix  $P$  such that  $A = P^{-1}JP$  and  $J$  is in canonical Jordan form.

The matrix  $P$  is the transition matrix to the basis from the theorem above. Notice that the total number of ones over the diagonal in  $J$  equals the difference between the total algebraic and geometric multiplicity of the eigenvalues. The ordering of the blocks in the matrix corresponds to the chosen ordering

<sup>5</sup>Camille Jordan was a famous French Mathematician working in Analysis and Algebra at the end of the 19th and the beginning of the 20th centuries.



`A.is_unitary()`, which for our cases returns True. Alternatively, we can assess the orthonormality of the matrix's columns (or rows) with respect to the standard Hermitian form on  $\mathbb{C}^n$  (for a  $n \times n$  matrix). This is analogous to evaluating orthogonality in real matrices. These arguments enable us to conclude that the matrices  $B$  is unitary, while the matrix  $C$  is not (although the scaled matrix  $\frac{1}{2}C$  is unitary). We will demonstrate the verification process for the matrix  $B$ , and leave the other case for practise.

For the matrix  $B$ , the column vectors are given by  $u_1 = (\frac{1+i}{2}, \frac{1-i}{2})^T$ ,  $u_2 = (\frac{\sqrt{2}}{2}, \frac{i\sqrt{2}}{2})^T$ . Thus

$$\begin{aligned} \|u_1\| &= \sqrt{\frac{1+i}{2} \cdot \frac{1-i}{2} + \frac{1-i}{2} \cdot \frac{1+i}{2}} = 1, \\ \|u_2\| &= \sqrt{\frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2}}{2} - i^2 \cdot \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2}}{2}} = 1, \\ \langle u_1, u_2 \rangle &= \frac{1+i}{2} \cdot \frac{\sqrt{2}}{2} - \frac{1-i}{2} \cdot \frac{i\sqrt{2}}{2} = 0. \end{aligned}$$

Hence  $B$  is unitary. Let us also proceed with Sage, by applying one of the methods described above. For example, executing the command

```
B = matrix(SR, [[(1+I)/2, sqrt(2)/2],
[(1-I)/2, sqrt(2)*I/2]])
B.is_unitary()
```

Sage returns True. □

Let  $V, W$  be two unitary spaces. Given a linear map  $\psi : V \rightarrow W$  its “adjoint” is the linear map  $\psi^* : W \rightarrow V$  defined by  $\langle \psi(u), w \rangle_W = \langle u, \psi^*(w) \rangle_V$ , for all  $u \in V$ , and  $w \in W$ . By the discussion in 3.4.5 we know that if  $A$  is the matrix of  $\psi$  in an orthonormal basis of  $V$ , then the conjugate transpose  $A^*$  is the matrix of  $\psi^*$ .

When  $f = f^*$ , or equivalently when  $A$  is Hermitian, i.e.,  $A = A^*$ , then  $f$  is called “self-adjoint”, see 3.4.6.<sup>16</sup> The following series of exercises will explore these objects, starting with the properties of the adjoint operator.

**3.D.22.** Consider two linear mappings  $\varphi, \psi : V \rightarrow V$  on a complex unitary space  $(V, \langle \cdot, \cdot \rangle)$ . Show that the adjoint operators of  $\varphi$  and  $\psi$  satisfy the following properties:

- $(c\varphi)^* = \bar{c}\varphi^*$ , for any  $c \in \mathbb{C}$ ;
- $(\varphi + \psi)^* = \varphi^* + \psi^*$ ;
- $(\varphi \circ \psi)^* = \psi^* \circ \varphi^*$ ;
- $(\varphi^*)^* = \varphi$ ;
- If  $\varphi$  is invertible, so is  $\varphi^*$  with  $(\varphi^*)^{-1} = (\varphi^{-1})^*$ . ○

<sup>16</sup>Self-adjoint operators and Hermitian matrices are fundamental in various scientific and engineering disciplines. In quantum mechanics, they represent observable quantities such as energy, momentum, and position with their real eigenvalues being of special importance. In numerical analysis, self-adjoint operators are used to solve eigenvalue problems, which are essential for simulations and modeling in engineering and physics. In computer science, Hermitian matrices find applications in machine learning, where they help in dimensionality reduction and data analysis by identifying important features in large datasets.

of the subspaces  $V_i$  in the direct sum. Thus, the uniqueness of the matrix  $J$  is true up to the ordering of the Jordan blocks. There is therefore freedom in the choice of the basis for such a Jordan canonical form.

**3.4.11. Remarks.** The existence of the Jordan canonical form is clear for the cases when all eigenvalues are either distinct or when the geometric and algebraic multiplicities of the eigenvalues are the same. In particular, this is the case for all unitary and self-adjoint mappings on unitary vector spaces, while the definition of normal mappings requires exactly this behavior. In particular, the Jordan canonical form of a mapping is diagonal if and only if the mapping is normal.

A consequence of the Jordan canonical form theorem is that for every linear mapping  $f$ , every eigenvalue of  $f$  uniquely determines an invariant subspace that corresponds to all Jordan blocks with this particular eigenvalue. We shall call this subspace the *root subspace* corresponding to the given eigenvalue.

We mention one useful corollary of the Jordan theorem (which is already used in the discussion about the behavior of Markov chains). Assume that the eigenvalues of our mapping  $f$  are all of absolute value less than one. Then repeated application of the linear mapping on every vector  $v \in V$  leads to a decrease of all coordinates of  $f^k(v)$  towards zero, without bounds.

Indeed, assume  $f$  has only one eigenvalue  $\lambda$  on all the complex space  $V$  and that  $f - \lambda \text{id}_V$  is cyclic (that is, we consider only one Jordan block separately). Let  $v_1, \dots, v_\ell$  be the corresponding basis. Then the theorem says that  $f(v_2) = \lambda v_2 + v_1$ ,  $f^2(v_2) = \lambda^2 v_2 + \lambda v_1 + \lambda v_1 = \lambda^2 v_2 + 2\lambda v_1$ , and similarly for other  $v_i$ 's and higher powers. In any case, the iteration of  $f$  results in higher and higher powers of  $\lambda$  for all non-zero components. The smallest of them can differ from the largest one only by less than the dimension of  $V$ . The coefficients are bounded too.

This proves the claim. The same argument can be used to prove that for the mapping with all eigenvalues with absolute value strictly greater than one leads to unbounded growth of all coordinates for the iterations  $f^k(v)$ .

The remainder of this part of the third chapter is devoted to the proof of the Jordan theorem and a few necessary lemmas. It is much more difficult than anything so far. The reader can skip it, until the beginning of the fifth part of this chapter in case of any problems with reading it.

**3.4.12. Root spaces.** We have already seen by explicit examples that the eigensubspaces completely describe geometric properties for some linear mappings only. Thus we now introduce a more subtle tool, the root subspaces.

**Definition.** A non-zero vector  $u \in V$  is called a *root vector* of the linear mapping  $\varphi : V \rightarrow V$ , if there exists an  $a \in \mathbb{K}$  and an integer  $k > 0$  such that  $(\varphi - a \text{id}_V)^k(u) = 0$ . This means that the  $k$ -th iteration of the given mapping sends  $u$



**3.D.23.** Let  $\varphi : \mathbb{C}^2 \rightarrow \mathbb{C}^2$  the linear mapping given by

$$\varphi \left( \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right) = \begin{pmatrix} iz_1 + 2z_2 \\ z_1 - iz_2 \end{pmatrix}.$$

Determine the matrix of  $\varphi$  with respect to the standard basis of  $\mathbb{C}^2$  and find its adjoint operator  $\varphi^*$ . Deduce that  $\varphi$  is *not* self-adjoint. Next solve the task in Sage.

**Solution.** Consider the standard orthonormal basis  $\underline{e} = \{e_1, e_2\}$  of  $\mathbb{C}^2$  (with respect to the standard Hermitian form  $\langle \cdot, \cdot \rangle$ ). Recall that if  $A = [\varphi]_{\underline{e}}$  is the matrix of  $\varphi$  with respect to  $\underline{e}$  and  $A = (a_{ij})$ , then we have  $a_{ij} = \langle \varphi(u_j), u_i \rangle$ . We compute  $\varphi(e_1) = (i, 1)^T$ ,  $\varphi(e_2) = (2, -i)^T$  and hence  $a_{11} = \langle \varphi(e_1), e_1 \rangle = i$ ,  $a_{12} = \langle \varphi(e_2), e_1 \rangle = 2$ ,  $a_{21} = \langle \varphi(e_1), e_2 \rangle = 1$  and  $a_{22} = \langle \varphi(e_2), e_2 \rangle = -i$ . Directly

$$A = [\varphi]_{\underline{e}} = \begin{pmatrix} \varphi(e_1) & \varphi(e_2) \end{pmatrix} = \begin{pmatrix} i & 2 \\ 1 & -i \end{pmatrix}.$$

Then, for the matrix  $A^* = \bar{A}^T$  we compute  $A^* = \begin{pmatrix} -i & 1 \\ 2 & i \end{pmatrix}$ .

Consequently, the adjoint of  $\varphi$  is given by

$$\varphi^* \left( \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right) = \begin{pmatrix} -iz_1 + z_2 \\ 2z_1 + iz_2 \end{pmatrix}.$$

To verify this result we will use defining equation of  $\varphi^*$ , that is,  $\langle \varphi(u), v \rangle = \langle u, \varphi^*(v) \rangle$ , for any two vectors  $u = (z_1, z_2)^T$  and  $v = (z_3, z_4)^T$  of  $\mathbb{C}^2$ . We compute

$$\begin{aligned} \left\langle \begin{pmatrix} iz_1 + 2z_2 \\ z_1 - iz_2 \end{pmatrix}, \begin{pmatrix} z_3 \\ z_4 \end{pmatrix} \right\rangle &= (iz_1 + 2z_2)\bar{z}_3 + (z_1 - iz_2)\bar{z}_4 \\ &= iz_1\bar{z}_3 + 2z_2\bar{z}_3 + z_1\bar{z}_4 - iz_2\bar{z}_4, \\ \left\langle \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \begin{pmatrix} -iz_3 + z_4 \\ 2z_3 + iz_4 \end{pmatrix} \right\rangle &= z_1\overline{(-iz_3 + z_4)} + z_2\overline{(2z_3 + iz_4)} \\ &= z_1(i\bar{z}_3 + \bar{z}_4) + z_2(2\bar{z}_3 - i\bar{z}_4) \\ &= iz_1\bar{z}_3 + z_1\bar{z}_4 + 2z_2\bar{z}_3 - iz_2\bar{z}_4. \end{aligned}$$

Finally, since  $A \neq A^*$  it follows that  $\varphi$  is not self-adjoint.

In Sage a solution goes as follows:

```
# Define the matrix A of the linear map phi
A = Matrix(QQbar, [[I, 2], [1, -I]])
# Display the matrix A
show("Matrix A of phi:")
show(A)
# Compute the adjoint of A
A_adjoint = A.H
# Display the adjoint matrix
show("Adjoint matrix A*:")
show(A_adjoint)
# Check if A is self-adjoint
is_self_adjoint = (A == A_adjoint)
# Display whether A is self-adjoint
show("Is A self-adjoint?")
show(is_self_adjoint)
```

□

**3.D.24. The numerical system QQbar in Sage.** Observe that the Sage solution presented in 3.D.23 uses the option QQbar to define the matrix A. This represents the numerical system of algebraic complex numbers. The field of algebraic

to zero. The set of all root vectors corresponding to a fixed scalar  $\lambda$  along with the zero vector is called the *root subspace* associated with the scalar  $\lambda \in \mathbb{K}$ . We denote it by  $\mathcal{R}_\lambda$ .

If  $u$  is a root vector and the integer  $k$  from the definition is chosen as the smallest possible one for  $u$ , then  $(\varphi - a \text{id}_V)^{k-1}(u)$  is an eigenvector with the eigenvalue  $a$ . Thus we have  $\mathcal{R}_\lambda = \{0\}$  for all scalars  $\lambda$  which are not in the spectrum of the mapping  $\varphi$ .

**Proposition.** Let  $\varphi : V \rightarrow V$  be a linear mapping. Then

- (1)  $\mathcal{R}_\lambda \subset V$  is a vector subspace for every  $\lambda \in \mathbb{K}$ ,
- (2) for every  $\lambda, \mu \in \mathbb{K}$ , the subspace  $\mathcal{R}_\lambda$  is invariant with respect to the linear mapping  $(\varphi - \mu \text{id}_V)$ . In particular  $\mathcal{R}_\lambda$  is invariant with respect to  $\varphi$ ,
- (3) if  $\mu \neq \lambda$ , then  $(\varphi - \mu \text{id}_V)|_{\mathcal{R}_\lambda}$  is invertible,
- (4) the mapping  $(\varphi - \lambda \text{id}_V)|_{\mathcal{R}_\lambda}$  is nilpotent.

**PROOF.** (1) Checking the properties of the vector subspace is easy and is left to the reader.

(2) Assume that  $(\varphi - \lambda \text{id}_V)^k(u) = 0$  and put  $v = (\varphi - \mu \text{id}_V)(u)$ . Then

$$\begin{aligned} (\varphi - \lambda \text{id}_V)^k(v) &= (\varphi - \lambda \text{id}_V)^k((\varphi - \lambda \text{id}_V) + (\lambda - \mu) \text{id}_V)(u) \\ &= (\varphi - \lambda \text{id}_V)^{k+1}(u) + (\lambda - \mu) \cdot (\varphi - \lambda \text{id}_V)^k(u) = 0 \end{aligned}$$

(3) If  $u \in \text{Ker}(\varphi - \mu \text{id}_V)|_{\mathcal{R}_\lambda}$ , then

$$\begin{aligned} (\varphi - \lambda \text{id}_V)(u) &= (\varphi - \mu \text{id}_V)(u) + (\mu - \lambda)u = (\mu - \lambda)u. \\ \text{This implies } 0 &= (\varphi - \lambda \text{id}_V)^k(u) = (\mu - \lambda)^k u \text{ and thus } \\ \text{also } u &= 0 \text{ for } \lambda \neq \mu. \end{aligned}$$

(4) Choose a basis  $e_1, \dots, e_p$  of the subspace  $\mathcal{R}_\lambda$ . By definition, there exist integers  $k_i$  such that  $(\varphi - \lambda \text{id}_V)^{k_i}(e_i) = 0$ . In particular, the entire mapping  $(\varphi - \lambda \text{id}_V)|_{\mathcal{R}_\lambda}$  must be nilpotent. □

**3.4.13. Quotient spaces.** Our next aim is to show that the dimension of the root spaces always equals the algebraic multiplicity of the corresponding eigenvalues. First, we introduce some general useful technical tools.



#### QUOTIENT SPACES

**Definition.** Let  $U \subset V$  be a vector subspace. Define an equivalence relation on the set of all vectors in  $V$  by  $v_1 \sim v_2$  if and only if  $v_1 - v_2 \in U$ . Axioms of equivalence are easy to check. The set  $V/U$  of the classes of this equivalence is equipped by the operations defined by using representatives. That is, for classes  $[u]$  and  $[v]$  determined by the vectors  $u$  and  $v$ , set  $[v] + [w] = [v + w]$ ,  $a[u] = [au]$ . This is a well defined vector space called the *quotient vector space* of the space  $V$  by the subspace  $U$ .

Check the correctness of the definition of the operations and verify all axioms of the vector space in detail!

The classes (vectors) in the quotient space  $V/U$  will often be denoted as formal sums of one representative with all

numbers, usually denoted by  $\bar{\mathbb{Q}}$ , is formed by adjoining the rational numbers  $\mathbb{Q}$  with the roots of all polynomial equations with rational coefficients. In other words, an algebraic number is a complex number that is a root of a non-zero polynomial equation with rational coefficients. Thus, in Sage QQbar is an extension of the numerical system QQ, enabling exact computations with algebraic numbers. However, the result would be the same if we used CC, instead of QQbar, as both numerical systems approximate complex numbers.

Note however that QQbar allows for exact algebraic computations, while CC uses floating-point arithmetic for complex numbers. In particular, when displaying complex matrices with QQbar, then each element is shown in its exact algebraic form, preserving the precision of computations. In contrast, CC displays complex matrices using floating-point approximations, which can introduce rounding errors but are generally more efficient for large-scale computations.

**3.D.25.** Use Sage to determine which of the following matrices is Hermitian:

$$A = \begin{pmatrix} \sqrt{2} & i & 1-i \\ -i & 10 & \sqrt{2}+i \\ 1+i & \sqrt{2}-i & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} a & i+b \\ -i+b & \sqrt{|a|} \end{pmatrix}, \text{ where } a, b \in \mathbb{R},$$

$$C = \begin{pmatrix} 1 & 8i & 1-i\sqrt{5} \\ -8i & 4z & 0 \\ 1+i\sqrt{5} & 0 & 1 \end{pmatrix}, \text{ where } z \in \mathbb{C} \text{ with } \text{Im}(z) = 0,$$

$$D = \begin{pmatrix} i\sqrt{2} & i & 1-i \\ -i & 0 & 4+i \\ 1+i & 4-i & 0 \end{pmatrix}.$$

**3.D.26.** Prove that with respect to the Frobenius inner product  $B(A, B) = \text{tr}(B^* A)$  on  $\text{Mat}_n(\mathbb{C})$  the following statement is true: Two matrices  $A, B \in \text{Mat}_n(\mathbb{C})$  satisfying  $B = U^* A U$  for some  $n \times n$  unitary matrix  $U$ , are such that  $\|A\|_B^2 = \|B\|_B^2$ .

**3.D.27.** Consider the linear map  $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$  defined by

$$\varphi \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) = \begin{pmatrix} \sqrt{2}x + y \\ x - y \\ 2y \end{pmatrix}.$$

Compute its adjoint operator  $\varphi^* : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ . Next examine the linear operator  $\psi : \mathbb{C}^3 \rightarrow \mathbb{C}^3$  whose matrix with respect to the standard basis of  $\mathbb{C}^3$  is given by

$$B = \begin{pmatrix} 10 & 5+5i & 3+2i \\ 5-5i & 5 & \sqrt{2}i \\ 3-2i & -\sqrt{2}i & 0 \end{pmatrix}.$$

vectors in the subspace  $U$ , for instance  $u + U \in V/U, u \in V$ . The class  $0 + U$  is the zero vector in  $V/U$ , i.e. the vector  $u \in V$  represents the zero element in  $V/U$  if and only if  $u \in U$ .

Trivial examples are  $V/\{0\} \cong V, V/V \cong \{0\}$ . Another example is the quotient space of the plane  $\mathbb{R}^2$  factored by any one-dimensional subspace (here, every one-dimensional subspace  $U \subset \mathbb{R}^2$  is a line passing through the origin). Then the equivalence classes are all the lines parallel to this line.

**Proposition.** Let  $U \subset V$  be a vector subspace and  $(u_1, \dots, u_n)$  be a basis of  $V$ , such that  $(u_1, \dots, u_k)$  is a basis of  $U$ . Then  $\dim V/U = n - k$  and the vectors

$$u_{k+1} + U, \dots, u_n + U$$

form a basis of  $V/U$ .

**PROOF.**  $V = \text{span}\{u_1, \dots, u_n\}$ , so  $V/U = \text{span}\{u_1 + U, \dots, u_n + U\}$ . But the first  $k$  generators are zero, thus  $V/U = \text{span}\{u_{k+1} + U, \dots, u_n + U\}$ . Assume that the linear combination  $a_{k+1}(u_{k+1} + U) + \dots + a_n(u_n + U) = (a_{k+1}u_{k+1} + \dots + a_nu_n) + U = 0 \in V/U$  vanishes. Equivalently, this linear combination of the vectors  $u_{k+1}, \dots, u_n$  belongs to the subspace  $U$ . Since  $U$  is generated by the remaining vectors in the basis of  $V$ , the latter linear combination is necessarily zero, and so all coefficients  $a_i$  are zero. This proves the linear independence of the generators of  $V/U$ .  $\square$

**3.4.14. Induced mappings on quotient spaces.** Assume that  $U \subset V$  is an invariant subspace with respect to linear mapping  $\varphi : V \rightarrow V$  and choose basis  $u_1, \dots, u_n$  of the space  $V$  such that the first  $k$  vectors of this basis is a basis of  $U$ . With



this basis,  $\varphi$  has block matrix  $A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$ . Then we can prove the following lemma:

**Lemma.** (1) the mapping  $\varphi$  induces a linear mapping  $\varphi_{V/U} : V/U \rightarrow V/U, \varphi_{V/U}(v+U) = \varphi(v)+U$  with the matrix  $D$  under the induced basis  $u_{k+1} + U, \dots, u_n + U$  on  $V/U$ ,  
 (2) the characteristic polynomial of  $\varphi_{V/U}$  divides the characteristic polynomial of  $\varphi$ .

**PROOF.** For  $v, w \in V, u \in U, a \in \mathbb{K}$  we have  $\varphi(v+u) \in \varphi(v)+U$  (because  $U$  is invariant),  $(\varphi(v)+U) + (\varphi(w)+U) = \varphi(v+w)+U$  and  $a(\varphi(v)+U) = a\varphi(v)+U = \varphi(av)+U$  (because  $\varphi$  is linear), thus the mapping  $\varphi_{V/U}$  is well-defined and linear. Moreover the very definition of the matrix of a mapping in a basis implies that the matrix of  $\varphi_{V/U}$  in the induced basis on  $V/U$  is exactly the matrix  $D$  (when counting the images of the basis elements the coefficients of the matrix  $C$  add only to the class  $U$ ).

The characteristic polynomial of the induced mapping  $\varphi_{V/U}$  is thus  $|D - \lambda E|$ , while characteristic polynomial of the original mapping  $\varphi$  is  $|A - \lambda E| = |B - \lambda E||D - \lambda E|$ .  $\square$

Demonstrate that this operator is self-adjoint. ○

**3.D.28.** Derive the adjoint of the trace  $\text{tr} : \text{Mat}_m(\mathbb{F}) \rightarrow \mathbb{F}$  with respect to the Frobenius inner product  $\mathcal{B}(A, B) = \text{tr}(B^* A)$ , for  $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ . ○

**3.D.29.** Let  $(V, \langle \cdot, \cdot \rangle)$  be an inner product space and let  $W \subset V$  be a subspace of  $W$ . Show that the orthogonal projection  $\text{proj}_W : V \rightarrow V$  on  $W$  is self-adjoint. ○



Recall that a linear endomorphism  $f : V \rightarrow V$  on a (finite-dimensional) vector space  $V$  is diagonalizable if and only if  $V$  admits a basis  $B$ , consisting of eigenvectors of  $f$ . When  $V$  is equipped with a scalar product  $\langle \cdot, \cdot \rangle$  and given an endomorphism  $f$ , it is natural to ask whether there exists an orthonormal basis of eigenvectors for  $f$ . If such a basis exists, the matrix  $[f]_B$  of  $f$  with respect to  $B$  is diagonal, and the same holds true for the matrix associated with the adjoint  $f^*$  of  $f$ . Since diagonal matrices commute, this implies  $f \circ f^* = f^* \circ f$ . Linear operators (or matrices) satisfying this relation are called “normal”. For instance, orthogonal or unitary matrices are normal, but there are many other types of normal matrices as well, see below.

It turns out that on a complex scalar vector space  $(V, \langle \cdot, \cdot \rangle)$  an endomorphism  $f : V \rightarrow V$  is normal if and only if  $V$  has an orthonormal basis consisting of eigenvectors of  $f$ . It is important to note that this result does not hold for real vector spaces. In the case of a real unitary space the existence of an orthonormal basis of eigenvectors for a linear operator  $f$  is equivalent to  $f$  being self-adjoint, i.e.,  $f = f^*$ .

**3.D.30.** Prove that orthogonal and unitary matrices are normal. Next demonstrate that not all normal matrices (or transformations) are orthogonal or unitary, by providing a counterexample, that is, a normal matrix that is neither orthogonal nor unitary. ○

Let  $\varphi : V \rightarrow V$  be a normal operator on a unitary space  $(V, \langle \cdot, \cdot \rangle)$ . By 3.4.8 and the proof given in the main theorem of this paragraph, we know that eigenvectors corresponding to distinct eigenvalues of  $\varphi$  are orthogonal with respect to  $\langle \cdot, \cdot \rangle$ . Let us illustrate this important situation by examples.

**3.D.31.** Suppose that  $\varphi : V \rightarrow V$  is a normal operator on a (finite-dimensional) unitary space  $(V, \langle \cdot, \cdot \rangle)$ , having two distinct eigenvalues, namely,  $\lambda_1 = 3$  and  $\lambda_2 = 5$ . Assume also that the corresponding eigenvectors are of unit length. Use the Pythagorean theorem to prove that there exist a vector  $u \in V$  with  $\|u\| = 2$  and  $\|\varphi(u)\| = 6$ , where  $\|\cdot\|$  is the norm induced by  $\langle \cdot, \cdot \rangle$ .

**Solution.** Suppose that  $v, w \in V$  are the eigenvectors of  $\varphi$  corresponding to  $\lambda_1$  and  $\lambda_2$ , that is  $\varphi(v) = 3v$  and  $\varphi(w) = 5w$ , respectively. By assumption  $\varphi$  is normal, hence eigenvectors corresponding to distinct eigenvalues are orthogonal,  $\langle v, w \rangle = 0$ . Then, for the vector  $u = v + w$  by the

**Corollary.** Let  $V$  be a vector space over  $\mathbb{K}$  of dimension  $n$  and let  $\varphi : V \rightarrow V$  be a linear mapping whose spectrum contains  $n$  elements (that is, all roots of the characteristic polynomial lie in  $\mathbb{K}$  and we count their multiplicities). Then there exists a sequence of invariant subspaces  $\{0\} = V_0 \subset V_1 \subset \dots \subset V_n = V$  with dimensions  $\dim V_i = i$ . Consider a basis  $u_1, \dots, u_n$  of the space  $V$  such that  $V_i = \text{span}\{u_1, \dots, u_i\}$ . In this basis, the matrix of the mapping  $\varphi$  is an upper triangular matrix:

$$\begin{pmatrix} \lambda_1 & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix},$$

with the spectrum  $\lambda_1, \dots, \lambda_n$  on the diagonal.

**PROOF.** The subspaces  $V_i$  are constructed inductively. Let  $\{\lambda_1, \dots, \lambda_n\}$  be the spectrum of the mapping  $\varphi$ . Thus the characteristic polynomial of the mapping  $\varphi$  is of the form  $\pm(\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$ . We choose  $V_0 = \{0\}$ ,  $V_1 = \text{span}\{u_1\}$ , where  $u_1$  is an eigenvector with eigenvalue  $\lambda_1$ . According to the previous theorem, the characteristic polynomial of the mapping  $\varphi_{V/V_1}$  is of the form  $\pm(\lambda - \lambda_2) \cdots (\lambda - \lambda_n)$ . Assume that we have already constructed linearly independent vectors  $u_1, \dots, u_k$  and invariant subspaces  $V_i = \text{span}\{u_1, \dots, u_i\}$ ,  $i = 1, \dots, k < n$  such that the characteristic polynomial of  $\varphi_{V/V_k}$  is of the form  $\pm(\lambda - \lambda_{k+1}) \cdots (\lambda - \lambda_n)$  and  $\varphi(u_i) \in (\lambda_i \cdot u_i + V_{i-1})$  for all  $i = 1, \dots, k$ .

We want to add one more vector  $u_{k+1}$  with analogous properties. There exists an eigenvector  $u_{k+1} + V_k \in V/V_k$  of the mapping  $\varphi_{V/V_k}$  with the eigenvalue  $\lambda_{k+1}$ . Consider the space  $V_{k+1} = \text{span}\{u_1, \dots, u_{k+1}\}$ . If the vector  $u_{k+1}$  is a linear combination of the vectors  $u_1, \dots, u_k$  then  $u_{k+1} + V_k$  would be the zero class in  $V/V_k$ . But this is not possible. Thus  $\dim V_{k+1} = k + 1$ . It remains to study the induced mapping  $\varphi_{V/V_{k+1}}$ . The characteristic polynomial of this mapping is of degree  $n - k - 1$  and divides the characteristic polynomial of the mapping  $\varphi$ . But completing the vectors  $u_1, \dots, u_{k+1}$  to the basis of  $V$  yields a block matrix of the mapping  $\varphi$  with an upper triangular submatrix  $B$  in the left upper corner and zero in the left lower corner. The diagonal elements are exactly the scalars  $\lambda_1, \dots, \lambda_{k+1}$ . Therefore the roots of the characteristic polynomial of the induced mapping have the required properties. □

**Remark.** If  $V$  decomposes into the direct sum of eigensubspaces for  $\varphi$ , the latter results do not say anything new. But their significance consists in the fact, that only the existence of  $\dim V$  roots of the characteristic polynomial (counting multiplicities) is assumed. This is ensured whenever the field  $\mathbb{K}$  is algebraically closed, for instance the complex numbers  $\mathbb{C}$ . As a direct consequence we see that the determinant and the trace of the mapping  $\varphi$  are always the product and the sum of the elements in the spectrum, respectively.

This can be also used for all real matrices. Just consider them to be complex, calculate the determinant or the trace as

Pythagorean theorem we have  $\|u\|^2 = \|v\|^2 + \|w\|^2$ , that is  $\|u\| = \sqrt{2}$ . On the other hand, again by the Pythagorean theorem we obtain

$$\begin{aligned} \|\varphi(u)\| &= \|\varphi(v+w)\| = \|\varphi(v) + \varphi(w)\| \\ &= \|3v + 5w\| = \sqrt{9 + 25} = 6. \end{aligned}$$

□

**3.D.32.** Consider the normal operator  $\varphi : \mathbb{C}^3 \rightarrow \mathbb{C}^3$  whose matrix with respect to the standard basis of  $\mathbb{C}^3$  is given by

$$A = \begin{pmatrix} 2 & 1+i & 0 \\ 1-i & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Use Sage to illustrate the orthogonality of eigenvectors corresponding to distinct eigenvalues of  $\varphi$ . Next present a formal solution. ○

**3.D.33. Skew-Hermitian matrices.** Recall that a complex matrix  $A = (a_{ij})$  is called skew-Hermitian if  $A = -A^*$ , or equivalently  $a_{ij} = -\bar{a}_{ji}$ . Prove that:

- (i) Every skew-Hermitian matrix  $A$  is normal.
- (ii) The eigenvalues of a skew-Hermitian matrix  $A$  are purely imaginary, that is, they are of the form  $i\mu$  with  $\mu \in \mathbb{R}$ .

**Solution.** (i) The first claim is obvious: If  $A^* = -A$ , then  $A^*A = -A^2 = AA^*$ . Hence  $A$  is normal.

(ii) Suppose that  $\lambda$  is an eigenvalue of  $A$  with corresponding eigenvector  $u$ , that is  $Au = \lambda u$ . It is sufficient to prove that  $\lambda = -\bar{\lambda}$ . First we see that

$$u^*Au = u^*(\lambda u) = \lambda u^*u = \lambda \|u\|^2,$$

and since  $\|u\| \in \mathbb{R}$  the r.h.s belongs to  $\mathbb{C}$ . Thus  $(u^*Au)^* = \bar{\lambda} \|u\|^2$ . Now, because  $\|u\| \in \mathbb{R}$ , taking conjugates we obtain

$$\begin{aligned} \bar{\lambda} \|u\|^2 &= \overline{\lambda \|u\|^2} = (u^*Au)^* = u^*A^*u = -u^*Au \\ &= -u^*\lambda u = -\lambda \|u\|^2. \end{aligned}$$

This final relation yields the desired  $\lambda = -\bar{\lambda}$ . □

**3.D.34.** Present an example of a normal matrix which is neither Hermitian nor skew-Hermitian. Moreover, use Sage to verify your answer. ○

Recall that a  $n \times n$  matrix  $A$  is diagonalizable if and only if it has  $n$  linearly independent eigenvectors, and in this case  $A$  is similar to the diagonal matrix  $D$  consisting of the eigenvalues of  $A$ , i.e.,  $P^{-1}AP = D$ . Here,  $P$  is the matrix obtained from the eigenvectors of  $A$  (as column vectors). A square matrix  $A$  is said to be orthogonally diagonalizable when an orthogonal matrix  $P$  can be found such that  $P^{-1}AP = P^TAP$  is diagonal.

To clarify the concept, we provide an example, and additional tasks on orthogonal diagonalization are discussed in Section F (see for example 3.F.46 for an implementation of orthogonal diagonalization using Sage).



the product or sum of eigenvalues and because both determinant and the trace are algebraic expressions in terms of the elements of the matrix, the results will be correct.

**3.4.15. Orthogonal triangulation.** If we are given a scalar product on a vector space  $V$  and  $U \subset V$  is a subspace, then clearly  $V/U \simeq U^\perp$  where  $v \in U^\perp$  is identified with  $v + U$ . Moreover, each class of the quotient space  $V/U$  contains exactly one vector from  $U^\perp$  (the difference of two such vector is in  $U \cap U^\perp$ ). We can exploit this observation in every inductive step of the proof of the theorem above. Choose the representative  $u_{k+1} \in V_k^\perp$  of the eigenvector of  $\varphi_{V/V_k}$ . This modification leads to the orthogonal basis with the properties required in the claim about triangulation in the corollary above. Therefore there exists such an orthonormal basis, and we arrive at a very important theorem:



SCHUR'S ORTHOGONAL TRIANGULATION THEOREM

**Theorem.** Let  $\varphi : V \rightarrow V$  be a linear mapping on a vector space with scalar product. Let there be  $m = \dim V$  eigenvalues, counting multiplicities. Then there exists an orthonormal basis of the space  $V$  such that the matrix of  $\varphi$  in this basis is upper triangular with eigenvalues  $\lambda_1, \dots, \lambda_m$  on the diagonal.

**3.4.16. Theorem.** Let  $\varphi : V \rightarrow V$  be a linear mapping and  $\lambda_1, \dots, \lambda_k$  be all distinct eigenvalues. Then the sum of the root spaces  $\mathcal{R}_{\lambda_1}, \dots, \mathcal{R}_{\lambda_k}$  is direct. Furthermore, for every eigenvalue  $\lambda$  the dimension of the subspace  $\mathcal{R}_\lambda$  equals the algebraic multiplicity of  $\lambda$ .

**PROOF.** We prove first the independence of nonzero vectors from different root spaces. We proceed by induction over the number  $k$  of root spaces. The claim is obvious if  $k = 1$ . Assume that the theorem holds for cases with less than  $k > 1$  root spaces and assume that vectors  $u_1 \in \mathcal{R}_{\lambda_1}, \dots, u_k \in \mathcal{R}_{\lambda_k}$  satisfy  $u_1 + \dots + u_k = 0$ . Then,  $(\varphi - \lambda_k \text{id}_V)^j(u_k) = 0$  for suitable  $j$ , and moreover all  $y_i = (\varphi - \lambda_k \text{id}_V)^j(u_i)$  are non-zero vectors in  $\mathcal{R}_{\lambda_i}$ ,  $i = 1, \dots, k-1$ , whenever  $u_i$  are non-zero by Proposition 3.4.12. But at the same time



$$y_1 + \dots + y_{k-1} = (\varphi - \lambda_k \cdot \text{id}_V)^j \left( \sum_{i=1}^k u_i \right) = 0$$

and, according to the inductive assumption, all  $y_i$  are zero. But then also all  $u_i$ ,  $1 \leq i < k$  must vanish and thus  $u_k = 0$ , too. This proves the first claim.

It remains consider the dimensions of the root spaces  $\mathcal{R}_\lambda$ . Consider an eigenvalue  $\lambda$  of  $\varphi$ , use the same notation  $\varphi$  for the restriction  $\varphi|_{\mathcal{R}_\lambda}$  and write  $\psi : V/\mathcal{R}_\lambda \rightarrow V/\mathcal{R}_\lambda$  for the mapping induced by  $\varphi$  on the quotient space.

Assume that the dimension  $\mathcal{R}_\lambda$  is strictly smaller than the algebraic multiplicity of the root  $\lambda$  of the characteristic polynomial. In view of lemma 3.4.14,  $\lambda$  is also an eigenvalue

**3.D.35. (Orthogonal diagonalization)** Consider the symmetric matrix  $A$  given below. Find a matrix  $P \in \text{Mat}_3(\mathbb{R})$  such that  $P^{-1}AP = P^T AP$  is diagonal:

$$A = \begin{pmatrix} 1 & 2 & 6 \\ 2 & 0 & 2 \\ 6 & 2 & 1 \end{pmatrix}.$$

A real  $n \times n$  symmetric matrix has real eigenvalues, and the same holds true for Hermitian matrices, see 3.4.7. However, the converse is not necessarily true, as demonstrated by the counterexample given by the matrix  $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ . Moreover, by the spectral decomposition theorem of self-adjoint operators (see the theorem in ??), the eigenvectors corresponding to distinct eigenvalues of a symmetric or a Hermitian matrix  $A$  are orthogonal. Leveraging these properties, one can derive the following fundamental result which is further motivated by the result in 3.D.35, mentioned above.

**3.D.36.** Prove that a real  $n \times n$  matrix  $A$  is orthogonally diagonalizable if and only if is symmetric, i.e.,  $A = A^T$ .

In a similar vein, a complex square matrix  $A \in \text{Mat}_n(\mathbb{C})$  is said to be “unitary diagonalizable”, if there exists a unitary matrix  $U$  such that  $U^*AU$  is diagonal. By the spectral theorem (see 3.4.7) we know that any Hermitian matrix  $A$  is unitary diagonalizable. Conversely, the following holds.



**3.D.37.** Suppose that  $A$  is a complex matrix with real eigenvalues, which can be diagonalizable by a unitary matrix  $U$ . Show that  $A$  is Hermitian.

**Solution.** By assumption, we have  $U^*AU = D$ , where  $D$  is the diagonal matrix having diagonal entries the eigenvalues of  $A$ . However, the hypothesis states that  $A$  has real eigenvalues, so  $D$  is a real diagonal matrix, which implies that  $D^* = D$ . Now, since  $U$  is unitary from the relation  $U^*AU = D$  we get  $A = (U^*)^{-1}DU^{-1} = UDU^*$ . Then we see that

$$A^* = (UDU^*)^* = UD^*U^* = UDU^* = A.$$

**3.D.38. (Unitary diagonalization)** Consider the matrices  $A, B$  given below. For each case demonstrate that the corresponding eigenvectors are orthogonal each other (with respect to the standard Hermitian form  $\langle \cdot, \cdot \rangle$  on  $\mathbb{C}^3$ ). Next demonstrate that both matrices are unitarily diagonalizable and illustrate the conditions necessary for unitary diagonalization:

$$A = \begin{pmatrix} -1 & 0 & i \\ 0 & 1 & 0 \\ -i & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & i \\ -i & 3 \end{pmatrix}.$$

**3.D.39.** Prove that a normal matrix  $A$  having all its eigenvalues purely imaginary is skew-Hermitian.

of the mapping  $\psi$ . Let  $(v + \mathcal{R}_\lambda) \in V/\mathcal{R}_\lambda$  be the corresponding eigenvector, that is,  $\psi(v + \mathcal{R}_\lambda) = \lambda(v + \mathcal{R}_\lambda)$ . Then  $v \notin \mathcal{R}_\lambda$  and  $\varphi(v) = \lambda v + w$  for suitable  $w \in \mathcal{R}_\lambda$ . Thus  $w = (\varphi - \lambda \text{id}_V)(v)$  and  $(\varphi - \lambda \text{id}_V)^j(w) = 0$  for suitable  $j$ . We conclude that  $(\varphi - \lambda \text{id}_V)^{j+1}(v) = 0$ , which contradicts the choice  $v \notin \mathcal{R}_\lambda$ .

It follows that the dimension of  $\mathcal{R}_\lambda$  equals the algebraic multiplicity of the root  $\lambda$  of the characteristic polynomial of the mapping  $\varphi : V \rightarrow V$ .  $\square$

Combining the latter theorem with the triangulation result from Corollary 3.4.14, we can formulate:

**Corollary.** Consider a linear mapping  $\varphi : V \rightarrow V$  on a vector space  $V$  over scalars  $\mathbb{K}$ , whose entire spectrum is in  $\mathbb{K}$ . Then  $V = \mathcal{R}_{\lambda_1} \oplus \dots \oplus \mathcal{R}_{\lambda_n}$  is the direct sum of the root subspaces. If we choose suitable bases for these subspaces, then under this basis  $\varphi$  has block-diagonal form with upper triangular matrices in the blocks and eigenvalues  $\lambda_i$  on the diagonal.



**3.4.17. Nilpotent and cyclic mappings.** Now almost everything is prepared for the discussion about canonical forms of matrices. It only remains to clear the relation between cyclic and nilpotent mappings and combine already proved results.

**Theorem.** Let  $\varphi : V \rightarrow V$  be a nilpotent linear mapping. Then there exists a decomposition of  $V$  into a direct sum of subspaces  $V = V_1 \oplus \dots \oplus V_k$  such that the restriction of  $\varphi$  to each summand  $V_i$  is cyclic.

**PROOF.** We provide a straightforward construction of a basis of the space  $V$  such that the action of the mapping  $\varphi$  on the basis vectors directly shows the decomposition into the cyclic mappings.

Let  $k$  be the degree of nilpotency of the mapping  $\varphi$  and write  $P_i = \text{Im}(\varphi^i)$ ,  $i = 0, \dots, k$ . Thus,

$$\{0\} = P_k \subset P_{k-1} \subset \dots \subset P_1 \subset P_0 = V.$$

Choose a basis  $e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1}$  of the space  $P_{k-1}$ , where  $p_{k-1} > 0$  is the dimension of  $P_{k-1}$ . By definition,  $P_{k-1} \subset \text{Ker } \varphi$ , i.e.  $\varphi(e_j^{k-1}) = 0$  for all  $j$ .

Assume that  $P_{k-1} \neq V$ . Since  $P_{k-1} = \varphi(P_{k-2})$ , there necessarily exist the vectors  $e_j^{k-2}$ ,  $j = 1, \dots, p_{k-1}$  in  $P_{k-2}$ , such that  $\varphi(e_j^{k-2}) = e_j^{k-1}$ . Assume

$$a_1 e_1^{k-1} + \dots + a_{p_{k-1}} e_{p_{k-1}}^{k-1} + b_1 e_1^{k-2} + \dots + b_{p_{k-1}} e_{p_{k-1}}^{k-2} = 0.$$

Applying  $\varphi$  on this linear combination yields  $b_1 e_1^{k-1} + \dots + b_{p_{k-1}} e_{p_{k-1}}^{k-1} = 0$ . This is a linear combination of independent vectors, therefore all  $b_j = 0$ . But then also  $a_j = 0$ . Thus the linear independence of all  $2p_{k-1}$  chosen vectors is established. Next, extend them to a basis

$$(1) \quad \begin{matrix} e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1} \\ e_1^{k-2}, \dots, e_{p_{k-1}}^{k-2}, e_{p_{k-1}+1}^{k-2}, \dots, e_{p_{k-2}}^{k-2} \end{matrix}$$

of the space  $P_{k-2}$ . The images of the added basis vectors are in  $P_{k-1}$ . Necessarily they must be linear combinations of



We have seen that working with diagonalizable matrices significantly simplifies many computations, such as matrix powers. When  $A$  is not diagonalizable, it is still useful to have a simplified form of  $A$  with respect to some ordered basis.

In fact, for any linear transformation  $f : V \rightarrow V$  on a finite-dimensional complex vector space  $V$ , there exists a particular simple matrix representation. If  $A$  is a complex square matrix, this means that we can find a matrix  $J$  such that  $A = PJP^{-1}$  for some invertible matrix  $P$ .

The matrix  $J$  is known as the Jordan normal form (or Jordan canonical form) and is typically represented as:

$$J = \begin{pmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_k \end{pmatrix}.$$

In this expression,  $J_i$  ( $i = 1, \dots, k$ ) are the so called Jordan blocks, which are upper triangular matrices whose form is explained below. The Jordan canonical form  $J$  will be a diagonal matrix if and only if  $A$  is diagonalizable, see 3.4.10.

Let us briefly outline the procedure for computing the Jordan canonical form of a given  $n \times n$  matrix  $A$ :

- Step 1: Compute the eigenvalues  $\lambda_1, \dots, \lambda_k$  of  $A$  and their algebraic multiplicities, which we denote by  $\alpha(\lambda_i)$ .
- Step 2: Compute the geometric multiplicity  $\gamma(\lambda_i)$  of each eigenvalue  $\lambda_i$  of  $A$ . Recall that  $\gamma(\lambda_i) = \dim \text{Ker}(A - \lambda_i E)$ , where  $E$  is the  $n \times n$  identity matrix (we have  $\gamma(\lambda_i) \leq \alpha(\lambda_i)$  for all  $i = 1, \dots, k$ ).
- Step 3: For each eigenvalue  $\lambda_i$  determine the root subspace (also known as generalized eigenspace)

$$\mathcal{R}_{\lambda_i} = \text{Ker}((A - \lambda_i E)^j),$$

where  $j = \alpha(\lambda_i)$  is the algebraic multiplicity of  $\lambda_i$ . Note the null spaces  $\text{Ker}((A - \lambda_i E)^\ell)$  for  $\ell < j = \alpha(\lambda_i)$  are included in  $\mathcal{R}_{\lambda_i}$ , for any eigenvalues  $\lambda_i$ , as  $\mathcal{R}_{\lambda_i}$  captures all the necessary vectors. In particular, the dimension of  $\mathcal{R}_{\lambda_i}$  coincides with the algebraic multiplicity  $j = \alpha(\lambda_i)$  of  $\lambda_i$ , for all  $i = 1, \dots, k$ .

- Step 4: Construct the Jordan block  $J_i \equiv J_{\lambda_i}$  for all  $i = 1, \dots, k$ , which has necessarily  $\lambda_i$  on the diagonal and 1s on the superdiagonal, while other elements are 0. To determine the size of  $J_i$  we start from the highest power  $j$  for which the dimension  $\dim \text{Ker}((A - \lambda_i E)^j)$  is non-zero (this highest power is typically the algebraic multiplicity, i.e.,  $j = \alpha(\lambda_i)$ ), and move downwards to  $j = 1$ . For example, a Jordan block of size  $m \times m$  corresponding to the eigenvalue  $\lambda_i$  has the form

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ 0 & 0 & \lambda_i & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_i \end{pmatrix} \in \text{Mat}_m(\mathbb{C}).$$

the basis elements  $e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1}$ . We can thus adjust the chosen vectors  $e_{p_{k-1}+1}^{k-2}, \dots, e_{p_{k-2}}^{k-2}$  by adding the appropriate linear combinations of the vectors  $e_1^{k-2}, \dots, e_{p_{k-1}}^{k-2}$  with the result that they are in the kernel of  $\varphi$ . Thus we may assume our choice in the scheme (1) has this property.

Assume that we have already constructed a basis of the subspace  $P_{k-\ell}$  such that we can directly arrange it into the scheme similar to (1)

$$\begin{matrix} e_1^{k-1}, \dots, e_{p_{k-1}}^{k-1} \\ e_1^{k-2}, \dots, e_{p_{k-1}}^{k-2}, e_{p_{k-1}+1}^{k-2}, \dots, e_{p_{k-2}}^{k-2} \\ e_1^{k-3}, \dots, e_{p_{k-1}}^{k-3}, e_{p_{k-1}+1}^{k-3}, \dots, e_{p_{k-2}}^{k-3}, e_{p_{k-2}+1}^{k-3}, \dots, e_{p_{k-3}}^{k-3} \\ \vdots \\ e_1^{k-\ell}, \dots, e_{p_{k-1}}^{k-\ell}, e_{p_{k-1}+1}^{k-\ell}, \dots, e_{p_{k-2}}^{k-\ell}, e_{p_{k-2}+1}^{k-\ell}, \dots, e_{p_{k-\ell}}^{k-\ell} \end{matrix}$$

where the value of the mapping  $\varphi$  on any basis vector is located above it. The value is zero if there is nothing above that basis vector.

If  $P_{k-\ell} \neq V$ , then again there must exist vectors  $e_1^{k-\ell-1}, \dots, e_{p_{k-\ell}}^{k-\ell-1}$  which map to  $e_1^{k-\ell}, \dots, e_{p_{k-\ell}}^{k-\ell}$ . We can extend them to a basis  $P_{k-\ell-1}$ , say, by the vectors

$$e_{p_{k-\ell}+1}^{k-\ell-1}, \dots, e_{p_{k-\ell-1}}^{k-\ell-1}.$$

Again, exactly as when adjusting (1) above, we choose the additional basis vectors from the kernel of  $\varphi$ . and analogically as before we verify that we indeed obtain a basis for  $P_{k-\ell-1}$ .

After  $k$  steps we obtain a basis for the whole  $V$ , which has the properties given for the basis of the subspace  $P_{k-\ell}$ . Individual columns of the resulting scheme then generate the subspaces  $V_i$ . Additionally we have found the bases of these subspaces which show that corresponding restrictions of  $\varphi$  are cyclic mappings.  $\square$

**3.4.18. Proof of the Jordan theorem.** Let  $\lambda_1, \dots, \lambda_k$  be all the distinct eigenvalues of the mapping  $\varphi$ . From the assumptions of the Jordan theorem it follows that  $V = \mathcal{R}_{\lambda_1} \oplus \dots \oplus \mathcal{R}_{\lambda_k}$ .

The mappings  $\varphi_i = (\varphi|_{\mathcal{R}_{\lambda_i}} - \lambda_i \text{id}_{\mathcal{R}_{\lambda_i}})$  are nilpotent and thus each of the root spaces is a direct sum

$$\mathcal{R}_{\lambda_i} = P_{1,\lambda_i} \oplus \dots \oplus P_{j_i,\lambda_i}$$

of spaces on which the restriction of the mapping  $\varphi - \lambda_i \text{id}_V$  is cyclic. Matrices of these restricted mappings on  $P_{r,s}$  are Jordan blocks corresponding to the zero eigenvalue, the restricted mapping  $\varphi|_{P_{r,s}}$  has thus for its matrix the Jordan block with the eigenvalue  $\lambda_i$ .

For the proof of Jordan theorem it remains to verify the claim about uniqueness (up to reordering the blocks). Because the diagonal values  $\lambda_i$  are given as roots of the characteristic polynomial, their uniqueness is immediate. The decomposition to root spaces is unique as well. Thus, without loss of generality we may assume that there is just one eigenvalue  $\lambda$  and we are going to express the dimensions of



Later we will see that each Jordan block corresponds to a size which represents the length of a so-called chain of generalized eigenvectors associated with  $\lambda_i$ . The increase in dimension between the kernel of  $(A - \lambda_i E)^{j-1}$  and  $(A - \lambda_i E)^j$  is the number of blocks of size  $j$  or more.

We finally form the matrix  $J$  as posed above, i.e.,  $J$  consists of the Jordan blocks  $J_i$  which all lie along the diagonal, while the blank off-diagonal blocks are all zero. Therefore, in a Jordan matrix  $J$  the only possibly non-zero entries are those on the diagonal, which can attain any complex value (including 0), and those on the superdiagonal, which are either 0 or 1.

**3.D.40.** Let  $A$  be a matrix with characteristic polynomial  $\chi_A(\lambda) = (\lambda - \sqrt{2})^3$ . Illustrate all possible configurations for the Jordan canonical form of  $A$ .

**Solution.** Since the characteristic polynomial is of degree three, the matrix  $A$  should be  $3 \times 3$ . Since  $\chi_A(\lambda) = (\lambda - \sqrt{2})^3$ , the unique eigenvalue  $\lambda = \sqrt{2}$  has algebraic multiplicity 3. Now, the geometric multiplicity  $\gamma(\lambda)$  of  $\lambda$  satisfies  $1 \leq \gamma(\lambda) \leq 3$ . Hence there are three possible configurations (up to the ordering of the Jordan blocks):

- If  $\gamma(\lambda) = 1$ , then the Jordan canonical form will consist of a single Jordan block of size 3:

$$J = \begin{pmatrix} \sqrt{2} & 1 & 0 \\ 0 & \sqrt{2} & 1 \\ 0 & 0 & \sqrt{2} \end{pmatrix}.$$

- If  $\gamma(\lambda) = 2$ , then the Jordan canonical form will consist of one Jordan block of size 2 and one Jordan block of size 1:

$$J = \left( \begin{array}{cc|c} \sqrt{2} & 1 & 0 \\ 0 & \sqrt{2} & 0 \\ \hline 0 & 0 & \sqrt{2} \end{array} \right).$$

- If  $\gamma(\lambda) = 3$ , then the Jordan canonical form will consist of three Jordan blocks of size 1 each (in this case there exist three linearly independent eigenvectors, hence  $A$  is diagonalizable and  $J$  is diagonal):

$$J = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}.$$

□

**3.D.41.** Which of the following matrices represents a Jordan canonical form? Explain your answer.

$$G_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, G_2 = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

$$G_3 = \begin{pmatrix} 5 & 1 \\ 0 & 1 \end{pmatrix}, G_4 = \begin{pmatrix} -2 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, G_5 = (\pi),$$

individual Jordan blocks using the ranks  $r_k$  of the mapping  $(\varphi - \lambda \text{id}_V)^k$ . This will show that the blocks are uniquely determined (up to their order). On the other hand, changing the order of the blocks corresponds to renumbering the vectors of basis, thus we can obtain them in any order.

If  $\psi$  is a cyclic operator on an  $m$ -dimensional space, then the defect of the iterated mapping  $\psi^k$  is  $k$  for  $0 \leq k \leq m$ , while the defect is  $m$  for all  $k \geq m$ . This implies that if our matrix  $J$  of the mapping  $\varphi$  on the  $n$ -dimensional space  $V$  (remind we assume  $V = \mathcal{R}_\lambda$ ) contains  $d_k$  Jordan blocks of the order  $k$ , then the defect  $D_\ell = n - r_\ell$  of the matrix  $(J - \lambda E)^\ell$  is

$$D_\ell = d_1 + 2d_2 + \cdots + \ell d_\ell + \ell d_{\ell+1} + \cdots.$$

Now, taking the combination  $2D_k - D_{k-1} - D_{k+1}$  we cancel all those terms in the latter expression which coincide for  $\ell = k - 1, k, k + 1$  and we are left with

$$2D_k - D_{k-1} - D_{k+1} = d_k.$$

Substituting for  $D_\ell$ 's, we finally arrive at

$$d_k = 2n - 2r_k - n + r_{k-1} - n + r_{k+1} = r_{k-1} - 2r_k + r_{k+1}.$$

This is the requested expression for the sizes of the Jordan blocks and the theorem is proved.

**3.4.19. Remarks.** The proof of the theorem about the existence of the Jordan canonical form was constructive, but it does not give an efficient algorithmic approach for the construction. Now we show how our results can be used for explicit computation of the basis in which the given mapping  $\varphi : V \rightarrow V$  has its matrix in the canonical Jordan form.<sup>6</sup>



- (1) Find the roots of the characteristic polynomial.
- (2) If there are less than  $n = \dim V$  roots (counting multiplicities), then there is no canonical form.
- (3) If there are  $n$  linearly independent eigenvectors, there is a basis of  $V$  composed of eigenvectors under which  $\varphi$  has diagonal matrix.
- (4) Let  $\lambda$  be the eigenvalue with geometric multiplicity strictly smaller than the algebraic multiplicity and  $v_1, \dots, v_k$  be the corresponding eigenvectors. They should be the vectors on the upper boundary of the scheme from the proof of the theorem 3.4.17. We need to complete the basis by application of iterations  $\varphi - \lambda \text{id}_V$ . By doing this we also find in which row the vectors should be located. Hence we find the linearly independent solutions  $w_i$  of the equations  $(\varphi - \lambda \text{id})(w_i) = v_i$  from the rows below it. Repeat the procedure iteratively (that is, for  $w_i$  and so on). In this way, we find the “chains” of basis vectors that give invariant subspaces, where  $\varphi - \lambda \text{id}$  is cyclic (the columns from the scheme in the proof).

<sup>6</sup>There is a beautiful purely algebraic approach to compute the Jordan canonical form efficiently, but it does not give any direct information about the right basis. This algebraic approach is based on polynomial matrices and Weierstrass divisors. We shall not go into details in this textbook.



$$G_6 = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

○

**3.D.42.** Find the Jordan canonical form of the matrix

$$A = \begin{pmatrix} 5 & 4 & 2 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{pmatrix}.$$

**Solution.** *Step 1:* The characteristic polynomial is given by

$$\begin{aligned} \chi_A(\lambda) &= \begin{vmatrix} 5-\lambda & 4 & 2 \\ 0 & 1-\lambda & -1 \\ -1 & -1 & 3-\lambda \end{vmatrix} = (5-\lambda) \begin{vmatrix} 1-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} \\ &\quad - 4 \begin{vmatrix} 0 & -1 \\ -1 & 3-\lambda \end{vmatrix} + 2 \begin{vmatrix} 0 & 1-\lambda \\ -1 & -1 \end{vmatrix} \\ &= -\lambda^3 + 9\lambda^2 - 24\lambda + 16. \end{aligned}$$

We see that 1 is a root of  $\chi_A(\lambda)$ . To find all roots we can apply the Horner's scheme:

$$\begin{array}{r|rrrr} 1 & -1 & 9 & -24 & 16 \\ & & -1 & 8 & -16 \\ \hline & -1 & 8 & -16 & 0 \end{array}$$

Thus,

$$\begin{aligned} \chi_A(\lambda) &= -\lambda^3 + 9\lambda^2 - 24\lambda + 16 \\ &= (\lambda - 1)(-\lambda^2 + 8\lambda - 16) = (\lambda - 1)(\lambda - 4)^2, \end{aligned}$$

and  $A$  has two eigenvalues:  $\lambda_1 = 1$  with  $\alpha(\lambda_1) = 1$  and  $\lambda_2 = 4$  with  $\alpha(\lambda_2) = 2$ .

*Step 2:* Let us now compute the geometric multiplicities. For  $\lambda_1 = 1$  we need to solve the matrix equation  $(A - E)u = 0$  for some vector  $u = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ . Since

$$A - E = \begin{pmatrix} 4 & 4 & 2 \\ 0 & 0 & -1 \\ -1 & -1 & 2 \end{pmatrix},$$

we obtain the linear system

$$\{4x_1 + 4x_2 + 2x_3 = 0, x_3 = 0, -x_1 - x_2 + 2x_3 = 0\}.$$

The solution space of this system is 1-dimensional:

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r \\ -r \\ 0 \end{pmatrix} : r \in \mathbb{R} \right\}.$$

Thus  $\gamma(\lambda_1) = \dim \text{Ker}(A - I) = 1$ , in particular we may assume that  $V_1$  is generated by the eigenvector  $u_1 = (1, -1, 0)^T$ . For the second eigenvalue  $\lambda_2 = 4$  we have the matrix equation  $(A - 4E)v = 0$  for some vector  $v = (y_1, y_2, y_3)^T \in \mathbb{R}^3$ . Since

$$A - 4E = \begin{pmatrix} 1 & 4 & 2 \\ 0 & -3 & -1 \\ -1 & -1 & -1 \end{pmatrix},$$

The procedure is practical for matrices when the multiplicities of the eigenvalues are small, or at least when the degrees of nilpotency are small. For instance, for the matrix

$$A = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

we obtain the two-dimensional subspace of eigenvectors

$$\text{span}\{(1, 0, 0)^T, (0, 1, 0)^T\},$$

but we still do not know, which of them are the “ends of the chains”. We need to solve the equations  $(A - 2E)x = (a, b, 0)^T$  for (yet unknown) constants  $a, b$ . This system is solvable if and only if  $a = b$ , and one of the possible solutions is  $x = (0, 0, 1)^T$ ,  $a = b = 1$ . The entire basis is then composed of  $(1, 1, 0)^T$ ,  $(0, 0, 1)^T$ ,  $(1, 0, 0)^T$ . Note that we have free choices on the way and thus there are many such bases.

### 5. Decompositions of the matrices and pseudoinversions

Previously we concentrated on the geometric description of the structure of a linear mapping. Now we translate our results into the language of matrix decomposition. This is an important topic for numerical methods and matrix calculus in general.

Even when computing effectively with real numbers we use decompositions into products. The simplest one is the unique expression of every real number in the form

$$a = \text{sgn}(a) \cdot |a|,$$

that is, as a product of the sign and the absolute value. Proceeding in the same way with complex numbers, we obtain their polar form. That is, we write  $z = (\cos \varphi + i \sin \varphi)|z|$ . Here the complex unit plays the role of the sign and the other factor is a non-negative real multiple.

In the following paragraphs we list briefly some useful decompositions for distinct types of matrices. Remind, we met suitable decompositions earlier, for instance for positive semidefinite matrices in paragraph 3.4.9 when finding the square roots. We shall start with similar simple examples.

**3.5.1. LU-decomposition.** In paragraphs 2.1.7 and 2.1.8 we transformed matrices over scalars from any field into row echelon form. For this we used elementary row transformations, based on successive multiplication of our matrix by invertible lower triangular matrices  $P_i$ . In this way we added multiples of the rows above the currently transformed one.

Sometimes we also interchanged the rows, which corresponded to multiplication by a *permutation matrix*. That is a square matrix in which all elements are zero except exactly one value 1 in each row and column. To imagine why, consider a matrix with just one non-zero element in the first column but not in the first row. When we used the backwards

we obtain the linear system

$$\{y_1 + 4y_2 + 2y_3 = 0, 3y_2 + y_3 = 0, y_1 + y_2 + y_3 = 0\}.$$

Using Sage via the cell

```
A1=matrix([[1, 4, 2], [0, -3, -1], [-1, -1, -1]])
show(A1)
show(A1.rref())
```

we obtain the following reduced row echelon form

$$A - 4E = \begin{pmatrix} 1 & 4 & 2 \\ 0 & -3 & -1 \\ -1 & -1 & -1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 2/3 \\ 0 & 1 & 1/3 \\ 0 & 0 & 0 \end{pmatrix},$$

which can be verified by hand in a few steps. Therefore, the solution space is given by

$$\left\{ \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3}s \\ -\frac{1}{3}s \\ s \end{pmatrix} : s \in \mathbb{R} \right\}.$$

Consequently, we may assume that the eigenspace  $V_4$  is generated by the eigenvector  $u_2 = (2, 1, -3)^T$  (this corresponds to the choice  $s = -3$ ). In particular,  $\gamma(\lambda_2) = 1$ .

*Step 3:* Since the eigenvalue  $\lambda_2 = 4$  satisfies  $\gamma(\lambda_2) = 1 < \alpha(\lambda_2) = 2$ , the matrix  $A$  is not diagonalizable, and we should proceed with Step 3. In particular, for  $\lambda_2 = 4$  we have  $j = \alpha(\lambda_2)$ , and this algebraic multiplicity coincides with the dimension of the generalized eigenspace  $\mathcal{R}_{\lambda_2} = \text{Ker}((A - 4E)^2)$ , i.e.,  $\dim \text{Ker}((A - 4E)^2) = 2$  (see below in 3.D.44 for a description of  $\mathcal{R}_{\lambda_2}$ ). Thus, according to our algorithm, for the eigenvalue  $\lambda_2 = 4$  exists a unique Jordan block of size  $2 \times 2$  or larger, since

$$\dim \text{Ker}((A - 4E)^2) - \dim \text{Ker}((A - 4E)) = 2 - 1 = 1.$$

Counting dimensions we see that this has necessarily the form  $J_2 = \begin{pmatrix} 4 & 1 \\ 0 & 4 \end{pmatrix}$ , since for  $\lambda_1 = 1$  the corresponding Jordan block  $J_1$  is just a scale,  $J_1 = (1)$ .

*Step 4:* We are now able to describe the Jordan canonical form:

$$J = (1) \oplus \begin{pmatrix} 4 & 1 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}. \quad \square$$

*Having the Jordan canonical form  $J$  of a squared matrix  $A$ , we can derive a similarity matrix  $P$  such that  $P^{-1}AP = J$ , or, equivalently,  $A = PJP^{-1}$ . Such a matrix  $P$  is composed of columns that are eigenvectors and generalized eigenvectors of  $A$ . To understand how  $P$  is constructed, recall that for an eigenvalue  $\lambda_i$  of  $A$ , a generalized eigenvector  $u$  is defined as a non-zero vector such that  $(A - \lambda_i E)^j u = 0$  for some integer  $j > 0$ . If  $j$  is the smallest positive integer with this property, the set*

$$\{(A - \lambda_i)^{j-1}u, (A - \lambda_i)^{j-2}u, \dots, (A - \lambda_i)u, u\}$$

*is known as a “(Jordan) chain of length  $j$ ”, see also the discussion in 3.4.19.*

*The structure of these chains reflects the sizes and arrangements of the Jordan blocks in the Jordan form  $J$ . In*

elimination to transform the matrix into the blockwise form

$$\begin{pmatrix} E_h & 0 \\ 0 & 0 \end{pmatrix}$$

(*remind  $E_h$  stays for the unit matrix of rank  $h$* ) then we potentially needed to interchange columns as well. This was achieved by multiplying by a permutation matrix from the right hand side.

For simplicity, assume we have a square matrix  $A$  of size  $m$  and that Gaussian elimination does not force a row interchange. Thus all matrices  $P_i$  can be lower triangular with ones on diagonal. Finally we note that inverses of such  $P_i$  are again lower triangular with ones on the diagonal (either remember the algorithm 2.1.10 or the formula in 2.2.11). We obtain

$$U = P \cdot A = P_k \cdots P_1 \cdot A$$

where  $U$  is an upper triangular matrix. Thus

$$A = L \cdot U$$

where  $L$  is lower triangular matrix with ones on diagonal and  $U$  is upper triangular. This decomposition is called *LU-decomposition* of the matrix  $A$ . We can also absorb the diagonal values of  $U$  into a diagonal matrix  $D$  and obtain the *LDU-decomposition* where both  $U$  and  $L$  have just ones along the diagonal,  $A = LDU$ .

For a general matrix  $A$ , we need to add the potential permutations of rows during Gaussian elimination. Then we obtain the general result. (Think why we can always put the necessary permutation matrices to the most left and most right positions!)

#### LU-DECOMPOSITION

Let  $A$  be any square matrix of size  $m$  over a field of scalars. Then we can find lower triangular matrix  $L$  with ones on its diagonal, upper triangular matrix  $U$  and permutation matrices  $P$  and  $Q$ , all of size  $m$ , such that

$$A = P \cdot L \cdot U \cdot Q.$$

**3.5.2. Remarks.** As one direct corollary of the Gaussian elimination we can observe that, up to a choice of suitable bases on the domain and codomain, every linear mapping  $f : V \rightarrow W$  is given by a matrix in block-diagonal form with unit matrix of the size equal to the dimension of the image of  $f$ , and with zero blocks all around. This can be reformulated as follows: every matrix  $A$  of the type  $m/n$  over a field of scalars  $\mathbb{K}$  can be decomposed into the product

$$A = P \cdot \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} \cdot Q,$$

where  $P$  and  $Q$  are suitable invertible matrices.

Previously (in 3.4.10) we discussed properties of linear mappings  $f : V \rightarrow V$  over complex vector spaces. We



particular, each chain corresponds to a block, and the number of vectors in the chain indicates the block's size. Now each column in  $P$  corresponds to a vector in a Jordan chain. For a Jordan block of size  $j$  associated with  $\lambda_i$ , the matrix  $P$  will include  $j$  columns from the Jordan chain of a generalized eigenvector associated with  $\lambda_i$ .

Let us illustrate these facts with examples and for further exercises on the Jordan canonical form see Section F.

**3.D.43.** Using the matrices  $G_1, \dots, G_6$  given in 3.D.41, derive the characteristic polynomial of the corresponding matrix  $A$  for those cases that represent a Jordan canonical form.  $\circ$

**3.D.44.** For the matrix  $A$  provided in 3.D.42, determine a similarity matrix  $P$  such that  $P^{-1}AP = J$ , where  $J$  is Jordan canonical form of  $A$ , as obtained in 3.D.42.

**Solution.** We know that  $\mathcal{R}_{\lambda_2} = \text{Ker}((A - 4E)^2)$  is 2-dimensional. This is because the homogeneous linear system induced by the matrix equation  $(A - 4E)^2w = 0$ , i.e.,

$$\begin{pmatrix} -1 & -10 & -4 \\ 1 & 10 & 4 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

has a 2-dimensional solution space, given by

$$\left\{ \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} -10t - 4s \\ t \\ s \end{pmatrix} : t, s \in \mathbb{R} \right\}.$$

To derive that matrix  $P$  we first need to solve the matrix equation  $(A - 4E)w = u_2$ , where  $v_2 = (2, 1, -3)^T$  is the eigenvector corresponding to  $\lambda_2 = 4$  (see 3.4.19). Then, the vector  $w$  should be an element of  $\mathcal{R}_{\lambda_2}$  and the set  $\{(A - 4E)w, w\} = \{u_2, w\}$  will give rise to a Jordan chain for  $w$ . Moreover, the columns of  $P$  will be the vectors  $u_1, u_2$  and  $w$ , i.e.,  $P = (u_1 \mid u_2 \mid w)$ .

Let us find  $w$ . The augmented matrix of the linear system induced by the equation  $(A - 4E)w = u_2$ , is given by

$$B := (A - 4E \mid u_2) = \left( \begin{array}{ccc|c} 1 & 4 & 2 & 2 \\ 0 & -3 & -1 & 1 \\ -1 & -1 & -1 & -3 \end{array} \right),$$

and we can apply elementary row operations to obtain the corresponding RREF. Successively one obtains

$$\begin{aligned} B &\xrightarrow{R_3 \rightarrow R_1 + R_3} \left( \begin{array}{ccc|c} 1 & 4 & 2 & 2 \\ 0 & -3 & -1 & 1 \\ 0 & 3 & 1 & -1 \end{array} \right) \xrightarrow{R_3 \rightarrow R_2 + R_3} \\ &\left( \begin{array}{ccc|c} 1 & 4 & 2 & 2 \\ 0 & -3 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right) \xrightarrow{R_2 \rightarrow -\frac{1}{3}R_2} \left( \begin{array}{ccc|c} 1 & 4 & 2 & 2 \\ 0 & 1 & 1/3 & -1/3 \\ 0 & 0 & 0 & 0 \end{array} \right) \\ &\xrightarrow{R_1 \rightarrow R_1 - 4R_2} \left( \begin{array}{ccc|c} 1 & 0 & 2/3 & 10/3 \\ 0 & 1 & 1/3 & -1/3 \\ 0 & 0 & 0 & 0 \end{array} \right). \end{aligned}$$

You can verify this expression in Sage by the `rref()`-method, as we learned in Chapter 2.

showed that every square matrix  $A$  of dimension  $m$  can be decomposed into the product

$$A = P \cdot J \cdot P^{-1},$$

where  $J$  is a block-diagonal with Jordan blocks associated with the eigenvalues of  $A$  on the diagonal. Indeed, this is just a reformulation of the Jordan theorem, because multiplying by the matrix  $P$  and by its inverse from the other side corresponds in this case just to the change of the basis on the vector space  $V$  (with transition matrix  $P$ ). The quoted theorem says that every mapping has Jordan canonical form in a suitable basis.

Analogously, when discussing the self-adjoint mappings we proved that for real symmetric matrices or for complex Hermitian matrices there exists a decomposition into the product

$$A = P \cdot D \cdot P^*,$$

where  $D$  is the diagonal matrix with all (always real) eigenvalues on the diagonal, counting multiplicities. Indeed, we proved that there is an orthonormal basis consisting of eigenvectors. Thus the transition matrix  $P$  reflecting the appropriate change of the basis must be orthogonal. In particular,  $P^{-1} = P^*$ .

For real orthogonal mappings we derived analogous expression as for the symmetric ones, i.e.  $A = P \cdot B \cdot P^*$ . But in this case the matrix  $B$  is block-diagonal with blocks of size two or one, expressing rotations, mirror symmetry and identities with respect to the corresponding subspaces.

**3.5.3. Singular decomposition theorem.** We return to general linear mappings  $f : V \rightarrow W$  between vector spaces (generally distinct). We assume that scalar products are defined on both spaces and we restrict ourselves to orthonormal bases only.



If we want a similar decomposition result as above, we must proceed in a more refined way than in the case of arbitrary bases. But the result is surprisingly similar and strong:

#### SINGULAR DECOMPOSITION

**Theorem.** Let  $A$  be a matrix of the type  $m/n$  over real or complex scalars. Then there exist square unitary matrices  $U$  and  $V$  of dimensions  $m$  and  $n$ , and a real diagonal matrix  $D$  with non-negative elements of dimension  $r$ ,  $r \leq \min\{m, n\}$ , such that

$$A = U S V^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

and  $r$  is the rank of the matrix  $AA^*$ .

The matrix  $S$  is determined uniquely up to the order of the diagonal elements  $s_i$  in  $D$ . Moreover,  $s_i$  are the square roots of the positive eigenvalues  $d_i$  of the matrix  $AA^*$ .

If  $A$  is a real matrix, then the matrices  $U$  and  $V$  are orthogonal.

```
A=matrix([[5, 4, 2], [0, 1, -1],
[-1, -1, 3]])
E=identity_matrix(3)
AA=A-4*E
print(bool(AA==
matrix([[1, 4, 2], [0, -3, -1],
[-1, -1, -1]])))
b=vector([2, 1, -3])
B=AA.augment(b)
B.rref()
```

Therefore, we obtain the system (we assume that  $w$  has the form  $w = (x, y, z)^T$ )

$$\left\{x + \frac{2}{3}z = \frac{10}{3}, y + \frac{1}{3}z = -\frac{1}{3}, z \in \mathbb{R}\right\}$$

with solution space

$$\left\{\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{2}{3}s + \frac{10}{3} \\ -\frac{1}{2}s - \frac{1}{3} \\ s \end{pmatrix} : s \in \mathbb{R}\right\}.$$

To obtain a representative we may set  $s = 0$  which gives the generalized eigenvector  $w = (10/3, -1/3, 0)^T$ . Let us use Sage to verify that  $w$  is indeed a generalized eigenvector. To do so, we can simply run the following code block:

```
w=vector([10/3, -1/3, 0])
A=matrix([[5, 4, 2], [0, 1, -1], [-1, -1, 3]])
E=identity_matrix(3)
BB=(A-4*E)^2
bool(BB*w==0)
```

Sage's output is True and we can present the matrix  $P$ :

$$P = (u_1 \mid u_2 \mid w) = \begin{pmatrix} 1 & 2 & \frac{10}{3} \\ -1 & 1 & -\frac{1}{3} \\ 0 & -3 & 0 \end{pmatrix}.$$

To verify the correctness of this result, first compute  $P^{-1}$  and next check that  $P^{-1}AP = J$ . For convenience, we will use Sage to perform these computations promptly.

```
A=matrix([[5, 4, 2], [0, 1, -1],
[-1, -1, 3]])
J=matrix([[1, 0, 0], [0, 4, 1],
[0, 0, 4]])
P=matrix([[1, 2, 10/3], [-1, 1, -1/3],
[0, -3, 0]])
show(A, P, J, P.inverse())
print(P.inverse()*A*P==J)
```

We mention that Sage provides a built-in method to obtain the Jordan canonical form of a matrix, using the command `A.jordan_form()`. For instance, for our matrix  $A$  give the cell

```
A=matrix([[5, 4, 2], [0, 1, -1],
[-1, -1, 3]])
J,P=A.jordan_form(transformation=True)
show(J, P)
```

**PROOF.** Assume first that  $m \leq n$ . Denote by  $\varphi : \mathbb{K}^n \rightarrow \mathbb{K}^m$  the mapping between real or complex spaces with standard scalar products, given by the matrix  $A$  in the standard bases.



We can reformulate the statement of the theorem as follows: there exists orthonormal bases on  $\mathbb{K}^n$  and  $\mathbb{K}^m$  in which the mapping  $\varphi$  is given by the matrix  $S$  from the statement of the theorem.

As noted before, the matrix  $A^*A$  is positive semidefinite. Therefore it has only real non-negative eigenvalues and there exists an orthonormal basis  $\underline{u}$  of  $\mathbb{K}^n$  in which the corresponding mapping  $\varphi^* \circ \varphi$  is given by a diagonal matrix with eigenvalues on the diagonal. In other words, there exists a unitary matrix  $V$  such that  $A^*A = VBV^*$  for a real diagonal matrix  $B$  with non-negative eigenvalues  $(d_1, d_2, \dots, d_r, 0, \dots, 0)$  on the diagonal,  $d_i \neq 0$  for all  $i = 1, \dots, r$ . Thus

$$B = V^*A^*AV = (AV)^*(AV).$$

This is equivalent to the claim that the first  $r$  columns of the matrix  $AV$  are orthogonal, while the remaining columns vanish because they have zero norm.

Next, we denote the first  $r$  columns of  $AV$  as  $v_1, \dots, v_r \in \mathbb{K}^m$ . Thus,  $\langle v_i, v_i \rangle = d_i$ ,  $i = 1, \dots, r$ , and the normalized vectors  $u_i = \frac{1}{\sqrt{d_i}}v_i$  form an orthonormal system of non-zero vectors. Extend them to an orthonormal basis  $\underline{u} = u_1, \dots, u_m$  for the entire  $\mathbb{K}^m$ . Expressing the original mapping  $\varphi$  in the bases  $\underline{u}$  of  $\mathbb{K}^n$  and  $\underline{u}$  of  $\mathbb{K}^m$ , yields the matrix  $\sqrt{B}$ . The transformations from the standard bases to the newly chosen ones correspond to the multiplication from the left by a unitary (orthogonal) matrix  $U$  and from the right by  $V^{-1} = V^*$ . This is the claim of the theorem.

If  $m > n$ , we can apply the previous part of the proof to the matrix  $A^*$  which implies the desired result.

All the previous steps in the proof are also valid in the real domain with real scalars.  $\square$

This proof of the theorem about singular decomposition is constructive and we can indeed use it for computing the unitary (orthogonal) matrices  $U$  and  $V$  and the non-zero diagonal elements of the matrix  $S$ .

The diagonal values of the matrix  $D$  from the previous theorem are called *singular values of the matrix  $A$* .

**3.5.4. Further comments.** When dealing with real scalars, the singular values of a linear mapping  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  have a simple geometric meaning:



Let  $K \subset \mathbb{R}^n$  be the unit ball in the standard scalar product. The image  $\varphi(K)$  is always an  $m$ -dimensional ellipsoid (possibly degenerate). The singular values of the matrix  $A$  are then the norms of the main half-axes. The theorem says further that the original ball allows an orthogonal set of diameters, whose images are exactly the half-axes of this ellipsoid.

For square matrices it can be seen that  $A$  is invertible if and only if all singular values are non-zero. The ratio of the

Sage's output has the form

$$\left( \begin{array}{ccc|ccc} 1 & 0 & 0 & & & \\ \hline 0 & 4 & 1 & & & \\ 0 & 0 & 4 & & & \end{array} \right), \quad \left( \begin{array}{ccc} 1 & \frac{1}{2} & 1 \\ -1 & \frac{1}{4} & 0 \\ 0 & -\frac{3}{4} & -\frac{1}{4} \end{array} \right).$$

This verifies the expression of the matrix  $J$  derived in 3.D.42 and presents a similarity matrix  $P$ , such that  $P^{-1}AP = J$ . However, be aware that  $P$  is not uniquely determined, which is why Sage's candidate for  $P$  may differ from ours.  $\square$

**3.D.45.** Consider the matrix  $A = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}$ .

(a) Show that  $A$  is not diagonalizable and find the associated Jordan canonical form. Moreover, find a similarity matrix  $P$  such that  $P^{-1}AP = J$  and verify the results using Sage.

(b) Compute the 4th power  $A^4$  of  $A$  using the Jordan canonical form  $J$  and the matrices  $P, P^{-1}$ , derived in (a).  $\circ$

**3.D.46.** Consider the matrix  $A = \begin{pmatrix} 5 & 4 & 2 & 1 \\ 0 & 5 & 1 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 5 \end{pmatrix}$ . De-

termine the eigenvectors and generalized eigenvectors of  $A$  and find generators of the subspaces  $\text{Ker}((A - 5E)^2)$ ,  $\text{Ker}((A - 5E)^3)$  and  $\text{Ker}((A - 5E)^4)$ . Next determine the Jordan canonical form  $J$  of  $A$  and present a matrix  $P$  such that  $P^{-1}AP = J$ . Moreover show that the generalized eigenspace  $\mathcal{R}_\lambda$  of the unique eigenvalue  $\lambda$  of  $A$  satisfies  $\mathcal{R}_\lambda = \text{Ker}((A - 5E)^4) \cong \mathbb{R}^4$ .  $\circ$

### E. Matrix decompositions



*Matrices with some special property, such as diagonal matrices, triangular matrices, or unitary matrices, are simpler to work with than general matrices. The final part of this chapter focuses on matrix decompositions, which involving techniques expressing a given matrix as a product of "simpler" matrices, as those mentioned above, see the paragraphs 3.5.1 – 3.5.6. Matrix decompositions are essential for tasks like solving linear systems of equations and performing many complex matrix operations.*

We begin with the LU-decomposition, which is essentially the matrix form of the Gauss elimination method, described in paragraphs 2.1.7-2.1.9. The LU-decomposition exists for squared matrices  $A$  which can be reduced to a row-echelon form through a series of row operations without requiring row exchanges. In such a situation we can arrive to a factorization of the form  $A = LU$ , where  $U$  is the upper triangular matrix whose diagonal entries are the pivots, and  $L = (\ell_{ij})$  is a lower triangular matrix having units in the diagonal and nonzero off-diagonal entries  $\ell_{ij}$  encoding the elementary row operations that we applied to find  $U$ . Thus,  $\ell_{ij}$  represents the times that we subtracted the row  $j$  from row  $i$ , during the specific step of the Gaussian elimination.

greatest to the smallest singular value is an important parameter for the robustness of many numerical computations with matrices, for instance the computation of the inverse matrix. Note that there are fast methods of computation (approximations) for eigenvalues. Thus the singular decomposition is a very effective tool to work with.

**3.5.5. Polar decomposition theorem.** The singular decomposition theorem is the starting point for many other useful tools. We present several direct corollaries (which by themselves are non-trivial and important).

The statement of the singular decomposition theorem saying that for any matrix  $A$ , real or complex,  $A = USW^*$  with  $S$  diagonal with non-negative real numbers on the diagonal and  $U$  and  $W$  unitary, can be rephrased as

$$A = USU^*UW^*$$

and let us denote  $P = USU^*$ ,  $V = UW^*$ . The first of the matrices,  $P$ , is Hermitian (in the real case, symmetric) and positive semidefinite (because  $P$  and  $S$  are matrices of the the same mapping in different orthonormal bases). At the same time,  $V$  is the product of two unitary matrices and thus again is unitary (in the real case orthogonal).

Next, assume that  $A = PV = QZ$  are two such decompositions of the matrix  $A$  into the product of a positive semidefinite Hermitian matrix and a unitary matrix. Clearly,  $A^* = WSU^*$ . Thus  $AA^* = USSU^* = P^2$ , and the matrix  $P$  is actually the square root of the easily computable Hermitian matrix  $AA^*$ . In particular, this proves that  $P$  is uniquely determined, cf. 3.4.9.

Further, assume that  $A$  is invertible. Then also  $P$  is invertible and  $Z = V = P^{-1}A$ .

We have derived a very useful analogy of the decomposition of a real number into a sign and the absolute value:

#### POLAR DECOMPOSITION

**Theorem.** Every square complex matrix  $A$  of dimension  $n$  can be expressed in the form  $A = PV$ , where  $P$  is a Hermitian positive semi-definite square matrix of the same dimension while  $V$  is unitary.

The matrix  $P = \sqrt{AA^*}$  is uniquely given, and if  $A$  is invertible, the decomposition is unique and  $V = (\sqrt{AA^*})^{-1}A$ .

If  $A$  is a matrix of real scalars, then  $P$  is symmetric and  $V$  is orthogonal.

If we apply the same theorem to  $A^*$  instead of  $A$ , we obtain the same result, but with the order of the Hermitian and unitary matrices is reversed. This means  $A = VP$  with  $V$  unitary and  $P = \sqrt{A^*A}$  positive semidefinite. The matrices in the corresponding right and left polar decompositions will in general be different.

**3.E.1.** Find a  $LU$ -decomposition of the matrix

$$A = \begin{pmatrix} -2 & 1 & 0 \\ -4 & 4 & 2 \\ -6 & 1 & -1 \end{pmatrix}.$$

**Solution.** The desired decomposition is given by

$$A = LU = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -1 & 1 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

To obtain to this result, let us denote by  $R_1, R_2, R_3$  the rows of  $A$ , during any step of the Gaussian elimination. Then  $U$  (the row-echelon form of  $A$ ) occurs by the following elementary row operations

$$\begin{pmatrix} -2 & 1 & 0 \\ -4 & 4 & 2 \\ -6 & 1 & -1 \end{pmatrix} \begin{matrix} R_2 - \textcircled{2}R_1 \\ R_3 - \textcircled{3}R_1 \\ \Rightarrow \end{matrix} \begin{pmatrix} -2 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & -2 & -1 \end{pmatrix} \\ R_3 - \textcircled{-1}R_2 \Rightarrow \begin{pmatrix} -2 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix} = U.$$

Recall from Chapter 2 (cf. 2.A.9) that in Sage one can compute a row echelon form of a matrix  $A$  using the `echelon_form()` command. However, Sage's output may differ from the expected result, and some caution is required when choosing the numerical system. For instance, consider the matrix  $A$  with entries over  $\mathbb{Z}$ . Executing in Sage the cell given below, we will obtain the matrix  $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ , which is clearly not the desired result.

```
A = matrix(ZZ, [[-2, 1, 0], [-4, 4, 2], [-6, 1, -1]])
A.echelon_form()
```

On the other hand, you can verify on your own that by replacing `ZZ` with `RR`, `RDF`, `QQbar` or the symbolic ring `SR`, the resulting matrix will be the  $3 \times 3$  identity matrix. Hence, in this case, Sage returns the reduced row-echelon form of  $A$ , which is not useful for our task.

Now, for the lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{pmatrix}$$

we see that the entries  $\ell_{21}, \ell_{31}, \ell_{32}$  are given by the circled numbers 2, 3 and  $-1$ , appearing in the row operations mentioned above. In 3.E.5 below, you will be asked to derive the expression of  $L$  in terms of elementary matrices. Finally, a verification in Sage can be done manually, as follows:

Actually, if  $A$  is invertible, it is easy to check, that the matrices in the left are polar decomposition coincide if and only if  $A$  is normal. Look at theorem 3.4.8 and verify it yourself.

In the complex case the analogy with the decomposition of numbers is even more entertaining. The positive semidefinite  $P$  again plays the role of the absolute value of the complex number. The unitary matrix  $V$  uniquely allows the expression as a sum  $V = \operatorname{re} V + i \operatorname{im} V$  with Hermitian real and imaginary parts and the property  $(\operatorname{re} V)^2 + (\operatorname{im} V)^2 = E$ . We obtain a full analogy for the polar form for the complex numbers (see the final remark and corollary in 3.4.8). But note that in the higher dimensional case, it is important in which order this “polar form” of matrix is written. It is possible in both ways, but the results differ in general.



**3.5.6. QR decomposition.** For many practical applications it is faster to use another decomposition of matrices, which is an analogy of the Schur orthogonal triangulation theorem:

QR DECOMPOSITION

**Theorem.** For every complex matrix  $A$  of the type  $m/n$  there exists a unitary matrix  $Q$  and an upper triangular matrix  $R$  such that  $A = QR$ .

If all the scalars are real, then both  $Q$  and  $R$  are real (i.e.  $Q$  orthogonal).

**PROOF.** In the geometric formulation we need to prove that for every mapping  $\varphi : \mathbb{K}^n \rightarrow \mathbb{K}^m$  with the matrix  $A$  in the standard bases we can choose a new orthonormal basis on  $\mathbb{K}^m$  for which  $\varphi$  has upper triangular matrix.

Consider the images  $\varphi(e_1), \dots, \varphi(e_n) \in \mathbb{K}^m$  of the vectors of the standard orthonormal basis of  $\mathbb{K}^n$ . Choose from them a maximal linearly independent system  $v_1, \dots, v_k$  in such a way that the removed dependent vectors are always a linear combination of the previous vectors. Extend it into a basis  $v_1, \dots, v_m$ . Let  $u_1, \dots, u_m$  be an orthonormal basis  $\mathbb{K}^m$  obtained by the Gram-Schmidt orthogonalization of this system of vectors.

For every  $e_i$ ,  $\varphi(e_i)$  is either one of  $v_j$ ,  $j \leq i$ , or it is a linear combination of  $v_1, \dots, v_{i-1}$ . Therefore in the expression of  $\varphi(e_i)$  in the basis  $\underline{u}$  only the vectors  $u_1, \dots, u_i$  appear. Thus, in the standard basis on  $\mathbb{K}^n$  and  $\underline{u}$  on  $\mathbb{K}^m$ , the mapping  $\varphi$  has an upper triangular matrix  $R$ . The change of the basis  $\underline{u}$  on  $\mathbb{R}^m$  corresponds to the multiplication by a unitary matrix  $Q^*$  from the left. That is,  $R = Q^*A$ , equivalently  $A = QR$ .

The last claim is clear from the construction.  $\square$

**3.5.7. Pseudoinversions.** Finally, we discuss an especially useful and important extension of the inversion concept, which is of great importance for numerical procedures and also in Statistics.



```
A = matrix(RR, [[-2, 1, 0], [-4, 4, 2],
[-6, 1, -1]])
L = matrix(RR, [[1, 0, 0], [2, 1, 0],
[3, -1, 1]])
U = matrix(RR, [[-2, 1, 0], [0, 2, 2],
[0, 0, 1]])
L*U==A
```

Sage returns True. Note in this block we could use RDF or SR instead of RR, or even omit specifying a numerical system altogether.  $\square$

**3.E.2. Permuted LU-decomposition via Sage.** If partial pivoting (row exchanges) is allowed, then for any squared matrix  $A$  we have a decomposition of the form  $PA = LU$ , where  $P$  is a permutation matrix. This is known as the *LU-decomposition with partial pivoting*, or *permuted LU-decomposition*. Note that Sage provides the command `A.LU()`, which corresponds to a built-in method to obtain the permuted LU-decomposition. In particular, for a  $m \times n$  matrix  $A$  this command returns a triple of matrices  $P$ ,  $L$  and  $U$ , such that  $A = PLU$ , satisfying the following properties:  $P$  is a  $m \times m$  permutation matrix,  $L$  is a lower triangular  $m \times m$  matrix with units in the diagonal, and  $U$  is an upper triangular  $m \times n$  matrix. Notice in this case  $U$  is *not* in general a row-echelon form of  $A$ .

For example, using the matrix  $A$  from 3.E.1 and running the cell

```
A = matrix([[ -2, 1, 0], [-4, 4, 2], [-6, 1, -1]])
A.LU()
```

we obtain the following answer:

```
(
[0 0 1] [ 1 0 0] [ -6 1 -1]
[0 1 0] [2/3 1 0] [ 0 10/3 8/3]
[1 0 0], [1/3 1/5 1], [ 0 0 -1/5]
)
```

Notice that the matrices  $L$  and  $U$  that Sage prints out here, differ from those obtained in 3.E.1.

*The LU-decomposition is an effective method for solving linear systems. Given that  $A = LU$  is the LU-factorization of a squared matrix  $A$ , the equation  $Ax = b$  can be written as  $LUx = b$ . Introducing  $Ux = y$ , this gives us  $Ly = b$ . Typically, we first solve  $Ly = b$  with respect to  $y$  and then use this result to solve  $Ux = y$  for  $x$ .*

**3.E.3.** Solve  $Ax = b$  using a LU-decomposition of the matrix  $A$ , where  $A$  and  $b$  are respectively given by

$$A = \begin{pmatrix} 3 & -6 & 2 & -2 \\ -3 & 4 & 1 & 0 \\ 6 & -4 & 0 & -4 \\ -9 & 6 & -2 & 12 \end{pmatrix}, \quad b = \begin{pmatrix} -7 \\ 4 \\ 6 \\ -3 \end{pmatrix}.$$

**Solution.** One can verify that

Technically, the following quite straightforward application of singular decompositions of matrices allows us to define the pseudoinverse. However, we should beware that the singular decomposition is not unique and thus we must verify that such a definition is consistent. We shall see that in the next theorem.

PSEUDOINVERSE MATRICES

Let  $A$  be a real or complex matrix of the type  $m/n$ . Let

$$A = USV^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

be its singular decomposition (in particular,  $D$  is invertible). The matrix

$$A^\dagger := VS^\dagger U^*, \quad S^\dagger = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

is called the *pseudoinverse matrix* of the matrix  $A$ .

In geometric terms, we may view the linear mapping  $\varphi$  given by the matrix  $A$  in the two special orthonormal basis, where  $\varphi$  has got the matrix  $S$  with non-negative diagonal entries. We take the inverse of the “invertible part” of  $\varphi$  and complete it trivially to the pseudoinverse<sup>7</sup> mapping  $\varphi^\dagger$ . The result is then viewed in the original basis and yields  $A^\dagger$ .

As the following theorem shows, the pseudoinverse is an important generalization of the notion of inverse matrix, together with direct applications. At the same time, property (3), together with property (2), verifies the appropriateness of the definition.

<sup>7</sup>This concept was introduced by Eliakim Hastings Moore, an American mathematician, around 1920. It was reinvented by Roger Penrose and others later. In the literature, it is often called the *Moore-Penrose pseudoinverse*. Roger Penrose is an extremely influential mathematical physicist and philosopher of science working in Oxford, known also for his many best-selling popular books such as *The Emperor’s New Mind: Concerning Computers, Minds, and The Laws of Physics* (1989); *Shadows of the Mind: A Search for the Missing Science of Consciousness* (1994); *The Road to Reality: A Complete Guide to the Laws of the Universe* (2004); *Cycles of Time: An Extraordinary New View of the Universe* (2010).

$$A = LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & -4 & 1 & 0 \\ -3 & 6 & -14/8 & 1 \end{pmatrix} \begin{pmatrix} 3 & -6 & 2 & -2 \\ 0 & -2 & 3 & -2 \\ 0 & 0 & 8 & -8 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$$

For example, use Sage via the block

```
A=matrix(RDF, [[3, -6, 2, -2], [-3, 4, 1, 0],
[6, -4, 0, -4], [-9, 6, -2, 12]])
L=matrix(RDF, [[1, 0, 0, 0], [-1, 1, 0, 0],
[2, -4, 1, 0], [-3, 6, -14/8, 1]])
U=matrix(RDF, [[3, -6, 2, -2], [0, -2, 3, -2],
[0, 0, 8, -8], [0, 0, 0, 4]])
L*U == A
```

Sage prints out True. Now we solve the equation  $Ly = b$ , which is given by

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & -4 & 1 & 0 \\ -3 & 6 & -14/8 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} -7 \\ 4 \\ 6 \\ -3 \end{pmatrix}.$$

From this we obtain

$$\begin{aligned} y_1 &= -7, & y_2 &= 4 + y_1 = -3, \\ y_3 &= 6 - 2y_1 + 4y_2 = 8, \\ y_4 &= -3 + 3y_1 - 6y_2 + \frac{14}{8}y_3 = 8. \end{aligned}$$

Next we solve  $Ux = y$ , which is given by

$$\begin{pmatrix} 3 & -6 & 2 & -2 \\ 0 & -2 & 3 & -2 \\ 0 & 0 & 8 & -8 \\ 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -7 \\ -3 \\ 8 \\ 8 \end{pmatrix}.$$

By back substitution we get the desired solution:

$$\begin{aligned} x_4 &= \frac{8}{4} = 2, & x_3 &= \frac{8 + 8x_4}{8} = 3, \\ x_2 &= \frac{-3 - 3x_3 + 2x_4}{-2} = \frac{-8}{-2} = 4, \\ x_1 &= \frac{-7 + 6x_2 - 2x_3 + 2x_4}{3} = \frac{15}{3} = 5. \end{aligned}$$

□

Given a  $n \times n$  matrix  $A$  with a LU-decomposition  $A = LU$ , the efficiency of solving the linear equation  $Ax = b$  can be analyzed by counting the number of multiplications and divisions required.<sup>17</sup> As we saw before, to solve  $Ax = b$  using LU-factorization, the process involves solving  $Ly = b$  and  $Ux = y$ . It can be proved that approximately  $n^2$  multiplications and divisions are needed for these steps. For instance, in the case of a  $4 \times 4$  matrix  $A$  as in 3.E.3, you can count exactly 16 multiplications and divisions for these steps. For instance, in the case of a  $4 \times 4$  matrix  $A$  as in 3.E.3, you can count exactly 16 multiplications and divisions, 6 for solving the equation  $Ly = b$  and 10 for solving the equation  $Ux = y$ .

<sup>17</sup>Additions are typically ignored in this analysis due to their lower computational “cost”.

PROPERTIES OF PSEUDOINVERSE MATRICES

**Theorem.** Let  $A$  be a real or complex matrix of the type  $m/n$  and let  $A^\dagger$  be its pseudoinverse. Then:

- (1) if  $A$  is invertible (necessarily square), then
 
$$A^\dagger = A^{-1},$$
- (2)  $A^\dagger A$  and  $AA^\dagger$  are Hermitian (in real case symmetric) and
 
$$AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger.$$
- (3) the pseudoinverse matrices  $A^\dagger$  are uniquely defined by the four properties from (2). Thus if some matrix  $B$  of the type  $n \times m$  has the properties that  $BA$  and  $AB$  are both Hermitian,  $ABA = A$  and  $BAB = B$ , then  $B = A^\dagger$ .
- (4) if  $A$  is a matrix of the system of linear equations  $Ax = b$  with  $b \in \mathbb{K}^m$ , then the vector  $y = A^\dagger b \in \mathbb{K}^n$  minimizes the norm  $\|Ax - b\|$  for all vectors  $x \in \mathbb{K}^n$ .
- (5) the system of linear equations  $Ax = b$  with  $b \in \mathbb{K}^m$  is solvable if and only if  $AA^\dagger b = b$ . In this case all solutions are given by the expression

$$x = A^\dagger b + (E - A^\dagger A)u,$$

where  $u \in \mathbb{K}^n$  is arbitrary.

PROOF. (1): If  $A$  is invertible, then the matrix  $S = U^*AV$  is also invertible and directly from the definition  $S^\dagger = S^{-1}$ . Consequently,  $A^\dagger A = AA^\dagger = E$ .

(2): Direct computation yields  $SS^\dagger S = S$  and  $S^\dagger SS^\dagger = S^\dagger$ , therefore

$$AA^\dagger A = USV^*V S^\dagger U^*USV^* = USS^\dagger SV^* = USV^* = A$$

and analogically for the second equation. Furthermore,

$$\begin{aligned} (AA^\dagger)^* &= (USS^\dagger U^*)^* = U(S^\dagger)^* S^* U^* \\ &= U(SS^\dagger)^* U^* = USS^\dagger U^* = AA^\dagger. \end{aligned}$$

It can be proved similarly that  $(A^\dagger A)^* = A^\dagger A$ .

(3) The claim can be proved by direct computation. Of course we can consider the matrices  $A, A^\dagger, B$  as the matrices of the mappings  $\varphi, \varphi^\dagger$ , and  $\psi$  in the standard bases on  $\mathbb{K}^n$  and  $\mathbb{K}^m$ , or any other pair of orthonormal bases. The requested equality is equivalent to the equality  $\varphi^\dagger = \psi$  independently of the choice of the bases. We choose a couple of orthogonal bases from the singular decomposition of  $A$ . Then the mapping  $\varphi$  has the matrix  $S$  from the definition of the pseudoinverse  $A^\dagger$ , so we write directly

$$A = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad A^\dagger = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

with the diagonal matrix  $D$  consisting of all non-zero singular values. We write again  $B$  for the matrix of  $\psi$  in these bases. Clearly  $B$  and  $A$  satisfy the assumptions of the claim (3). Thus

$$A^\dagger A = \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}, \quad ABA = A$$



On the other hand, performing row reduction directly on the augmented matrix  $(A \mid b)$  to transform it to  $(U \mid y)$  requires about  $n^3/3$  arithmetic operations (as multiplications and divisions), which is significantly more time-consuming. Even worse, computing the inverse of  $A$  demands about  $n^3$  operations, and multiplying  $b$  by  $A^{-1}$  to obtain  $x = A^{-1}b$  requires another  $n^2$  operations.

**3.E.4. How are determinants of large matrices computed?** Typically, determinants of large matrices are not calculated using cofactor expansion, but rather through matrix decomposition. For instance, with an  $LU$ -decomposition of a squared matrix  $A$ , the determinant of  $A$  satisfies the relation  $\det(A) = \det(L)\det(U)$ . Since both  $L, U$  are triangular matrices, the previous relation implies that  $\det(A)$  equals the product of the elements on the diagonals of  $L$  and  $U$ .

**3.E.5.** Consider the  $LU$ -decomposition of the matrix  $A$  described in 3.E.1.

- i) Express the lower-triangular matrix  $L$  in terms of elementary matrices and compute its inverse.
- ii) Compute the inverse of the matrix  $U$  using Gauss elimination.
- iii) Calculate both the determinant and the inverse matrix of  $A$ , using its  $LU$ -decomposition as derived above.

○

Let us now focus on the singular value decomposition, which is possibly the most important matrix decomposition, with a wide range of remarkable applications. According to the singular value decomposition (in short SVD), given a matrix  $A \in \text{Mat}_{m,n}(\mathbb{F})$  where  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{F} = \mathbb{C}$ , we can find unitary matrices  $U \in \text{Mat}_m(\mathbb{F})$  and  $V \in \text{Mat}_n(\mathbb{F})$ , and a diagonal matrix  $\Sigma \in \text{Mat}_{m,n}(\mathbb{F})$  with non-negative entries such that  $A = U\Sigma V^*$ , see 3.5.3.

In such a decomposition the diagonal matrix  $\Sigma$  consists of the singular values of  $A$ , which are the non-negative square roots of the eigenvalues of  $A^*A$  (or equivalently of  $AA^*$ ). They are typically denoted by  $\sigma_1, \dots, \sigma_{\min\{m,n\}}$ , and we order them so that  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min\{m,n\}}$ .

In fact, there are exactly  $\text{rank}(A)$  non-zero singular values of  $A$ , as  $\text{rank}(U\Sigma V^*) = \text{rank}(\Sigma)$ , and the rank of a diagonal matrix coincides with the number of non-zero diagonal elements. On the other hand, the columns of  $V$  are (normalized) eigenvectors of  $A^*A$ , while the columns of  $U$  are (normalized) eigenvectors of  $AA^*$ .

and we obtain

$$A^\dagger = A^\dagger A B A A^\dagger = \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} B \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Consequently

$$B = \begin{pmatrix} D^{-1} & P \\ Q & R \end{pmatrix}$$

for suitable matrices  $P, Q$  and  $R$ . Next,

$$BA = \begin{pmatrix} D^{-1} & P \\ Q & R \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} E & 0 \\ QD & 0 \end{pmatrix}$$

is Hermitian. Thus  $QD = 0$  which implies  $Q = 0$  (the matrix  $D$  is diagonal and invertible). Analogously, the assumption that  $AB$  is Hermitian implies that  $P$  is zero. Finally, we compute

$$B = BAB = \begin{pmatrix} D^{-1} & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} D^{-1} & 0 \\ 0 & R \end{pmatrix}.$$

On the right side in the right-lower corner there is zero, and thus also  $R = 0$  and the claim is proved.

(4): Consider the mapping  $\varphi : \mathbb{K}^n \rightarrow \mathbb{K}^m, x \mapsto Ax$ , and direct sums  $\mathbb{K}^n = (\text{Ker } \varphi)^\perp \oplus \text{Ker } \varphi, \mathbb{K}^m = \text{Im } \varphi \oplus (\text{Im } \varphi)^\perp$  of the orthogonal complements. The restricted mapping  $\tilde{\varphi} := \varphi|_{(\text{Ker } \varphi)^\perp} : (\text{Ker } \varphi)^\perp \rightarrow \text{Im } \varphi$  is a linear isomorphism. If we choose suitable orthonormal bases on  $(\text{Ker } \varphi)^\perp$  and  $\text{Im } \varphi$  and extend them to orthonormal bases on whole spaces, the mapping  $\varphi$  will have matrix  $S$  and  $\tilde{\varphi}$  the matrix  $D$  from the theorem about the singular decomposition. In the next section, we shall discuss in detail that for any given  $b \in \mathbb{K}^m$ , there is the unique vector which minimizes the distance  $\|b - z\|$  among all  $z \in \text{Im } \varphi$  (in analytic geometry we shall say that the point  $z$  realises the distance of  $b$  from the affine subspace  $\text{Im } \varphi$ ), see 4.1.13). The properties of the norm proved in theorem 3.4.3 directly imply that this is exactly the component  $z = b_1$  of the decomposition  $b = b_1 + b_2, b_1 \in \text{Im } \varphi, b_2 \in (\text{Im } \varphi)^\perp$ .

Now, in our choice of bases, the mapping  $\varphi^\dagger$  is given by the matrix  $S^\dagger$  from the singular decomposition theorem. In particular,  $\varphi^\dagger(\text{Im } \varphi) = (\text{Ker } \varphi)^\perp, D^{-1}$  is the matrix of the restriction  $\varphi^\dagger|_{\text{Im } \varphi}$ , and  $\varphi^\dagger|_{(\text{Im } \varphi)^\perp}$  is zero. Indeed,

$$\varphi \circ \varphi^\dagger(b) = \varphi(\varphi^\dagger(z)) = z$$

and the proof is finished.

(5) Evidently, the equality  $Ax = b$ , with  $x \in \mathbb{K}^n$  fixed, implies

$$b = Ax = AA^\dagger Ax = AA^\dagger b.$$

Thus the condition is necessary. On the other hand, if this condition holds, then the choice  $x = A^\dagger b + (E - A^\dagger A)u$  as in (5) implies

$$Ax = A(A^\dagger b + (E - A^\dagger A)u) = b + (A - AA^\dagger A)u = b.$$

The rank of the matrix  $E - A^\dagger A$  gives the correct size of the image of the corresponding mapping according to the Kronecker-Capelli theorem (cf. 2.3.5) about the solution of the system of linear equations, and thus we obtain all solutions in this way.  $\square$



**3.E.6.** How many (non-zero) singular values has the matrix  $A = \begin{pmatrix} 4 & 0 \\ 3 & 5 \end{pmatrix}$ ? Describe them explicitly. ○

**3.E.7.** For the matrix  $A$  given below, derive the matrix  $\Sigma$ , as described above:

$$A = \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

There is a simple way to perform the SVD-decomposition, which is described as follows: First compute the matrices  $\Sigma$  and  $V$  as suggested above. To compute the unitary matrix  $U$ , which forms the last piece of the SVD-decomposition, we rely on the relation  $AV = U\Sigma$ . To simplify the description, suppose that  $\text{rank}(A) = r$  and express  $V, U$  in terms of columns, as  $V = (v_1 \mid v_2 \mid \cdots \mid v_n)$  and  $U = (u_1 \mid u_2 \mid \cdots \mid u_m)$  respectively. Then we have

$$\begin{aligned} AV &= (Av_1 \mid Av_2 \mid \cdots \mid Av_n) \\ &= (\sigma_1 u_1 \mid \cdots \mid \sigma_r u_r \mid 0 \mid \cdots \mid 0) = U\Sigma. \end{aligned}$$

Hence we will have  $u_k = \frac{1}{\sigma_k} Av_k$ , for all  $k = 1, \dots, r$ . To determine the remaining  $m - r$  columns of  $U$  it suffices to extend the set  $\{u_1, \dots, u_r\}$  to an orthonormal basis of  $\mathbb{F}^m$ , which usually can be done via the Gram-Schmidt procedure.

**3.E.8.** Describe the singular value decompositions of the matrices given in 3.E.6 and 3.E.7, respectively.

**Solution.** (a) Let us first consider the  $2 \times 2$  matrix  $A$ , where  $\Sigma = \text{diag}(2\sqrt{10}, \sqrt{10})$ . Let us derive the eigenvectors corresponding to the eigenvalues  $\lambda_1 = 40$  and  $\lambda_2 = 10$  of  $A^T A$ , to construct the matrix  $V$ . The solution space of the linear system induced by the matrix equation  $(A^T A - 40E)v$  has the form  $\{(x_1, x_2)^T = (t, t)^T : t \in \mathbb{R}\}$ . Hence, eigenvectors of  $A^T A$  corresponding to  $\lambda_1 = 40$  are scalar multiples of the vector  $(1, 1)^T$ . This vector has norm  $\sqrt{2}$  and we obtain the normalized eigenvector  $v_1 = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T$ . For the second eigenvalue we see that the linear system induced by the matrix equation  $(A^T A - 10E)v$  has solution space  $\{(x_1, x_2)^T = (-s, s)^T : s \in \mathbb{R}\}$ . Consequently, eigenvectors of  $A^T A$  corresponding to  $\lambda_2$  are scalar multiples of  $(-1, 1)^T$ . The norm of this vector is also  $\sqrt{2}$ , and we obtain the normalized eigenvector  $v_2 = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T$ . Therefore,

$$V = (v_1 \mid v_2) = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix},$$

**Remark.** Notice that the last computation in the proof verifies that  $(E - A^\dagger A)$  is the matrix of the projection of  $\mathbb{R}^n$  onto the subspaces of all solutions of the homogenous system  $Ax = 0$ .

It can be also shown that the matrix  $A^\dagger$  minimizes the square of the norm of the expression

$$AA^\dagger - E$$

that is, the sum of squares of all elements of the given matrix.

The claim (4) of the theorem can be also interpreted as follows.  $AA^\dagger$  is the matrix of the orthogonal projection from the vector space  $\mathbb{R}^m$ , onto the subspace generated by the columns of the matrix  $A$  ( $m$  is the number of the rows of the matrix  $A$ ). This interpretation has a strong meaning for matrices having more rows than columns. Moreover, for matrices  $A$  whose columns are independent vectors, the expression  $(A^T A)^{-1} A^T$  makes sense and it is not hard to verify that this matrix satisfies all the properties from (1) and (2) from the previous theorem. Thus it is the pseudoinverse  $A^\dagger$  of the matrix  $A$ .

**3.5.8. Linear regression.** The approximation property (4)



from the previous theorem is very useful in the cases where we are to find as good an approximation as possible for the (non-existent) solution of a given system  $Ax = b$ , where  $A$  is a real matrix of the type  $m/n$  and  $m > n$ .

For instance, an experiment gives many measured real values  $b_j$ ,  $j = 1, \dots, m$ . We want to find a linear combination of only a few fixed functions  $f_i$ ,  $i = 1, \dots, n$  which approximates the values  $b_j$  as good as possible. The actual values of the fixed functions at the relevant points  $y_j \in \mathbb{R}$  define the matrix  $a_{ij} = f_i(y_j)$ . The columns of the matrix are given by values of the individual functions  $f_i$  at the considered points. The goal is to determine the coefficients  $x_i \in \mathbb{R}$  so that the sum of the squares of the deviations from the actual values

$$\sum_{j=1}^m (b_j - \sum_{i=1}^n x_i f_i(y_j))^2 = \sum_{j=1}^m (b_j - \sum_{i=1}^n a_{ij} x_i)^2$$

is minimized. By the previous theorem, the optimal coefficients are  $A^\dagger b$ .

As an example, consider just three functions  $f_0(y) = 1$ ,  $f_1(y) = y$ ,  $f_2(y) = y^2$ . Assume that the “measured values” of their unknown combination  $g(y) = x_0 + x_1 y + x_2 y^2$  in integral values for  $y$  between 1 and 5 are  $b^T = (1, 12, 6, 27, 33)$ . (This vector  $b$  arose by computing the values  $1 + y + y^2$  at the given points adjusted by random integral values in the range  $\pm 10$ .) This leads in our case to the matrix  $A = (b_{ji})$

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 9 & 16 & 25 \end{pmatrix}.$$

and it is easy to verify that  $V$  is unitary. Next, for the matrix  $U = (u_1 \mid u_2)$  we have

$$u_1 = \frac{1}{\sigma_1} Av_1 = \frac{1}{2\sqrt{10}} \begin{pmatrix} 4 & 0 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{5}}{5} \\ \frac{2\sqrt{5}}{5} \end{pmatrix},$$

$$u_2 = \frac{1}{\sigma_2} Av_2 = \frac{1}{\sqrt{10}} \begin{pmatrix} 4 & 0 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} -\frac{2\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} \end{pmatrix}.$$

A direct computations show that  $u_1, u_2$  are orthonormal, and hence the matrix  $U$  is given by

$$U = (u_1 \mid u_2) = \begin{pmatrix} \frac{\sqrt{5}}{5} & -\frac{2\sqrt{5}}{5} \\ \frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{pmatrix},$$

such that  $A = U\Sigma V^T$ . For a quick confirmation, you may use Sage:

```
A=matrix([[4, 0], [3, 5]])
show("\n\nThe matrix A is given by A=", A)
S=diagonal_matrix([2*sqrt(10), sqrt(10)])
show("\n\nThe matrix S is given by S=", S)
V=matrix([[sqrt(2)/2, -sqrt(2)/2],
[sqrt(2)/2, sqrt(2)/2]])
show("\n\nThe matrix V is given by V=", V)
print(V.is_unitary())
U=matrix([[sqrt(5)/5, -2*sqrt(5)/5],
[2*sqrt(5)/5, sqrt(5)/5]])
show("\n\nThe matrix U is given by U=", U)
print(U.is_unitary())
print(A*V==U*S)
```

(b) In this case we have seen that  $\Sigma = \text{diag}(1, 1/2, 0)$ . For the eigenvalues  $\lambda_1 = 1, \lambda_2 = 1/4$  and  $\lambda_3 = 0$  of  $A^T A$  we compute the normalized eigenvectors  $v_1 = (1, 0, 0)^T, v_2 = (0, 0, 1)^T$  and  $v_3 = (0, 1, 0)^T$ . This can be confirmed in Sage by the following cell:

```
A=matrix([[0, 0, -1/2], [-1, 0, 0],
[0, 0, 0]])
AT=A.T; a= AT*A
show(a.eigenvalues())
S=diagonal_matrix([1, 1/2, 0])
show("\n\nThe matrix S is given by S=", S)
aein=a.eigenvectors_right()
show(aein)
```

which also prints the expression of the matrix  $\Sigma$  (denoted inside the cell by  $S$ ). Thus  $V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ . Next, for the

matrix  $U = (u_1 \mid u_2 \mid u_3)$  we can compute the first two columns by the rule  $u_k = \frac{1}{\sigma_k} Av_k$ . This gives

$$u_1 = \frac{1}{\sigma_1} Av_1 = \frac{1}{1} \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix},$$

$$u_2 = \frac{1}{\sigma_2} Av_2 = \frac{1}{1/2} \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}.$$

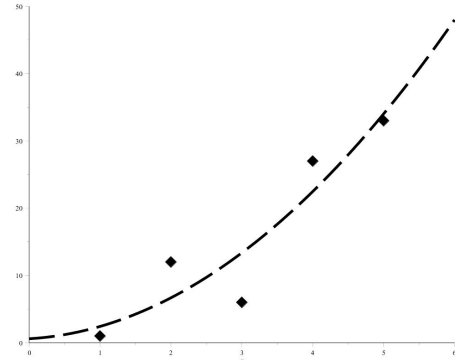
The requested optimal coefficients for the combination are

$$x = A^\dagger \cdot b$$

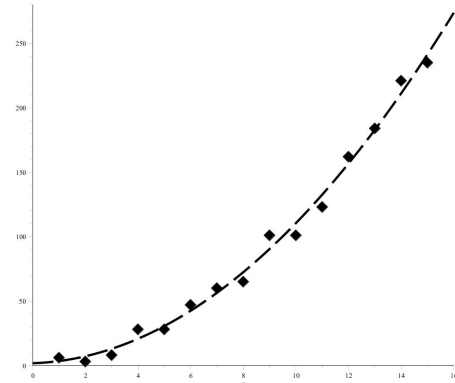
$$= \begin{pmatrix} \frac{9}{5} & 0 & -\frac{4}{5} & -\frac{3}{5} & \frac{3}{5} \\ -\frac{37}{35} & \frac{23}{70} & \frac{6}{7} & \frac{37}{70} & -\frac{23}{35} \\ \frac{1}{7} & -\frac{1}{14} & -\frac{1}{7} & -\frac{1}{14} & \frac{1}{7} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 12 \\ 6 \\ 27 \\ 33 \end{pmatrix}$$

$$\simeq \begin{pmatrix} 0.600 \\ 0.614 \\ 1.214 \end{pmatrix}.$$

The resulting approximation can be seen in the picture, where the given values  $b$  are shown by the diamonds, while the dashed curve stays for the resulting approximation  $g(y) = x_1 + x_2 y + x_3 y^2$ .



The computation was produced in Maple and taking 15 values  $y_i = 1 + i + i^2$ , with a random vector of deviations from the same range added produced the following picture:



It remains to extend  $\{u_1, u_2\}$  to an orthonormal basis of  $\mathbb{R}^3$ . Thus, we need to find a unit vector which is orthogonal to both  $u_1, u_2$ . Such a vector is the vector  $u_3 = (0, 0, -1)^T$  which is in fact the cross product on  $\mathbb{R}^3$  of  $u_1, u_2$ :

$$u_1 \times u_2 = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{vmatrix} = 0 \cdot \vec{i} - 0 \cdot \vec{j} + (-1) \cdot \vec{k}.$$

This gives the matrix

$$U = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

such that  $A = U\Sigma V^T$ , i.e.,

$$A = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

If you would like to verify this computation using Sage, you can continue by typing the following code in the previous block:

```
V=matrix([[1, 0, 0], [0, 0, 1], [0, 1, 0]])
show("\nThe matrix V is given by V=", V)
print(V.is_unitary())
U=matrix([[0, -1, 0], [-1, 0, 0], [0, 0, -1]])
show("\nThe matrix U is given by U=", U)
print(U.is_unitary())
A==U*S*(V.T)
```

In this way Sage confirms that both the matrices  $U$  and  $V$  are unitary, and moreover the relation  $A = U\Sigma V^T$ .

Note that Sage offers a built-in method to compute the singular value decomposition of a matrix, but it is only available for matrices over the numerical systems CDF or RDF. For our matrix  $A$ , for example, one can type

```
A=matrix(RDF, [[0, 0, -1/2], [-1, 0, 0], [0, 0, 0]])
show(A.SVD())
```

Note that Sage's output differs from our result only in terms of signs and column permutations. This variation is expected, as in SVD decomposition, the singular vectors (eigenvectors of  $U$  and  $V$ ) may differ by signs or orderings. Hence, be aware that different software implementations or methods might produce different but equally valid SVD results.  $\square$

*Using the components of the SVD-decomposition of a matrix  $A$  we can construct the polar decomposition, which is another significant matrix factorization, see 3.5.5. This describes a complex matrix  $A$  as a product of a unitary matrix  $U_p$  and a positive semi-definite matrix  $P$ . If  $A$  is real then  $U_p$  is an orthogonal matrix. Recall that saying  $P$  is positive semi-definite we mean that  $P$  has non-negative eigenvalues. If  $A = U\Sigma V^*$  is the SVD decomposition of  $A$ , then the corresponding polar decomposition  $A = PU_p$  occurs easily by setting  $U_p = UV^*$  and  $P = U\Sigma U^*$ , respectively.*

**3.E.9.** For the matrix  $A$  given in 3.E.7 compute its polar decomposition  $A = PU_p$  and use Sage to verify your result. Is

this decomposition unique? Next show that the matrix  $P$  is Hermitian and positive semi-definite.  $\circ$

Let  $A = U\Sigma V^*$  be the SVD-decomposition of a  $m \times n$  matrix  $A$  of rank  $r$ , where  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$  is the  $m \times n$  diagonal matrix with the singular values of  $A$ . The pseudo-inverse of  $A$  (also known as the Moore-Penrose-inverse) is defined by

$$A^\dagger = V\Sigma^\dagger U^*,$$

where  $\Sigma^\dagger = \text{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}, 0, \dots, 0\right)$ .

This matrix functions as a generalized inverse for matrices that are singular or non-square. As such, it extends the concept of an inverse matrix in several ways, see 3.5.7. Notably, while not every matrix has an inverse, every matrix, including non-square ones, has a pseudo-inverse.

**3.E.10.** For the matrices  $A$  given in 3.E.6 and 3.E.7 compute the corresponding pseudo-inverses.

**Solution.** (a) For the matrix  $A$  presented in 3.E.6 we obtain

$$\Sigma^\dagger = \begin{pmatrix} \frac{1}{2\sqrt{10}} & 0 \\ 0 & \frac{1}{\sqrt{10}} \end{pmatrix}.$$

Thus, the explicit form of the pseudo-inverse  $A^\dagger = V\Sigma^\dagger U^T$  of the matrix  $A$  is given by

$$\begin{aligned} A^\dagger &= \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2\sqrt{10}} & 0 \\ 0 & \frac{1}{\sqrt{10}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \\ -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} & 0 \\ -\frac{3}{20} & \frac{1}{5} \end{pmatrix}. \end{aligned}$$

To confirm our result we need to verify the identity  $AA^\dagger A = A$ , which is direct:

$$\begin{aligned} AA^\dagger A &= \begin{pmatrix} 4 & 0 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} \frac{1}{4} & 0 \\ -\frac{3}{20} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 3 & 5 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 3 & 5 \end{pmatrix} = A. \end{aligned}$$

In fact, observe that  $\det(A) = 20 \neq 0$ . Hence  $A$  is non-singular and has a unique inverse that should coincide with  $A^\dagger$ . This provides another confirmation:

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} 5 & 0 \\ -3 & 4 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & 0 \\ -\frac{3}{20} & \frac{1}{5} \end{pmatrix} = A^\dagger.$$

(b) For the matrix  $A$  presented in 3.E.7 we obtain

$$\Sigma^\dagger = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and hence the corresponding pseudo-inverse has the form

$$\begin{aligned} A^\dagger &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ -2 & 0 & 0 \end{pmatrix}. \end{aligned}$$


Note that  $\det(A) = 0$  hence this matrix does not have an inverse. To confirm the relation  $AA^\dagger A = A$  (or  $A^\dagger AA^\dagger = A^\dagger$ ), you can use Sage as follows;

```
A=matrix([[0, 0, -1/2],
[-1, 0, 0], [0, 0, 0]])
Aps=matrix([[0, -1, 0],
[0, 0, 0], [-2, 0, 0]])
print(A*Aps*A==A)
```

Or you can use the following routine, which relies on Scipy (or Numpy).

```
def pseudo_inverse(mat) :
    from scipy import linalg
    return matrix(linalg.pinv(mat))
Ap=pseudo_inverse(A)
show(Ap)
```

Note that one should add this cell in the previous block.  $\square$

 Let us conclude our description of matrix factorizations with the  $QR$ -decomposition, which expresses a complex matrix as a product of a unitary matrix  $Q$  and an upper-triangular matrix  $R$ , see 3.5.6. If for example  $A \in \text{Mat}_{m,n}(\mathbb{C})$  comes with  $n$  linearly independent columns, then using the Gram-Schmidt orthogonalization we can obtain the factorization  $A = QR$ , where  $Q$  is of type  $m \times n$  with orthonormal columns, and  $R$  is an invertible upper triangular matrix with positive diagonal entries.

Finding the  $QR$ -decomposition only requires simple algebraic operations and it can be useful to summarize the steps, for a matrix  $A \in \text{Mat}_{m,n}(\mathbb{C})$ , as above:

- Step 1 Via the Gram-Schmidt procedure find an orthogonal basis of the column space  $C(A)$  of  $A$ .
- Step 2 Normalize the vectors of this basis to get the columns of  $Q$ .
- Step 3 As the columns of  $Q$  are orthonormal with respect to the dot product, we have  $Q^T Q = E$ , where  $E$  is the  $n \times n$  identity matrix. Hence, the upper triangular matrix  $R$  is given by the formula  $R = Q^T A$ . Alternatively, if  $u_1, \dots, u_n$  are the linearly independent columns of  $A$ , and  $q_1, \dots, q_n$  are the orthonormal vectors obtained by the Gram-Schmidt procedure, then the entries  $r_{ij}$  of  $R$  are given by  $r_{ij} = \langle u_j, q_i \rangle$  for any  $1 \leq i, j \leq n$ , such that  $r_{ij} = 0$  for all  $i > j$ . Note that both  $Q, R$  are unique.

**3.E.11.** Present the  $QR$ -decomposition of the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

**Solution.** The column of  $A$  are the vectors  $E_1, E_2, E_3$  presented in 2.C.50. Recall that there we constructed the orthonormal basis  $\{\hat{w}_i : i = 1, 2, 3\}$  of the 3-dimensional subspace  $W = \text{span}_{\mathbb{R}}\{E_1, E_2, E_3\}$ . Obviously, this coincides with the column space of  $A$ , i.e.,  $W = C(A)$ . Thus the  $3 \times 4$

matrix  $Q$  is given by  $Q = (\hat{w}_1 \mid \hat{w}_2 \mid \hat{w}_3)$ , or in other words

$$Q = \begin{pmatrix} 1/2 & \sqrt{3}/6 & \sqrt{6}/6 \\ 1/2 & \sqrt{3}/6 & \sqrt{6}/6 \\ 1/2 & \sqrt{3}/6 & -\sqrt{6}/3 \\ 1/2 & -\sqrt{3}/2 & 0 \end{pmatrix}.$$

We can verify that  $Q^T Q = E$  in Sage, but again you should be careful in the following sense. If we use the number system QQbar to define the matrix  $Q$ , then the cell below gives True, which is the correct result.

```
Q=matrix(QQbar,
[[1/2, sqrt(3)/6, sqrt(6)/6],
[1/2, sqrt(3)/6, sqrt(6)/6],
[1/2, sqrt(3)/6, -sqrt(6)/3],
[1/2, -sqrt(3)/2, 0]])
Q.transpose()*Q == matrix.identity(3)
```

However, if we configure the matrix  $Q$ , using one of the number systems RR, CC, or CDF, Sage will return False. Can you explain the possible reasons why this might occur?

Next, the matrix  $R$  occurs by the multiplication  $Q^T A$ , which gives

$$R = \begin{pmatrix} 2 & 3/2 & 1 \\ 0 & \sqrt{3}/2 & \sqrt{3}/3 \\ 0 & 0 & \sqrt{6}/3 \end{pmatrix}.$$

Note that  $\det(R) = \sqrt{2} \neq 0$ , hence  $R$  is invertible, as it should be.

Sage provides a built-in method to obtain the  $QR$ -decomposition of a matrix  $A$ , based on the command `A.QR()`. Hence, you can quickly verify the result, just by typing

```
A=matrix(QQbar, [[1, 1, 1], [1, 1, 1],
[1, 1, 0], [1, 0, 0]])
Q, R = A.QR()
```

In case you like to confirm the expression of  $R$  in a “manual” way, you may use the following cell:

```
Q=matrix(QQbar,
[[1/2, sqrt(3)/6, sqrt(6)/6],
[1/2, sqrt(3)/6, sqrt(6)/6],
[1/2, sqrt(3)/6, -sqrt(6)/3],
[1/2, -sqrt(3)/2, 0]])
A = matrix(RR, [[1, 1, 1], [1, 1, 1],
[1, 1, 0], [1, 0, 0]])
R=(Q.T)*A; R
```

□

*The  $QR$ -decomposition of a matrix  $A$  is very useful when a system of linear equations  $Ax = b$  which has no solution is given, but an approximation as good as possible is needed.*



*In other words, the main point is to minimize  $\|Ax - b\|$ . According to the Pythagorean theorem, one has*

$$\|Ax - b\|^2 = \|Ax - b_{\parallel}\|^2 + \|b_{\perp}\|^2,$$

where  $b$  is decomposed into  $b_{\parallel}$  and into  $b_{\perp}$ . The first component belongs to the range of the linear transformation  $A$ , and the second one is perpendicular to this range. Now, the projection on the range of  $A$  can be written in the form  $QQ^T$  for a suitable matrix  $Q$ , which one obtains by the Gram-Schmidt orthonormalisation of the columns of  $A$ . Then, it follows that

$$Ax - b_{\parallel} = Q(Q^T Ax - Q^T b).$$

The system in the parentheses has a solution, for which  $\|Ax - b_{\parallel}\| = \|b_{\perp}\|$ , and this is the minimal value. Furthermore, the matrix  $R := Q^T A$  is upper triangular, therefore the approximate solution can be easily obtained.

**3.E.12.** Find an approximate solution of the following system of equations:

$$\{x_1 + 2x_2 = 1, 2x_1 + 4x_2 = 4\}.$$

**Solution.** We may express the given system in the matrix form  $Ax = b$ , where

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

Evidently, this system has no solution. Let us orthonormalize the columns of  $A$ : Take the first of them and divide it by its norm. This yields the first vector of the orthonormal basis  $\frac{1}{\sqrt{5}}(1, 2)^T$ . The second column vector of  $A$  is twice the first and thus we deduce that  $Q = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ . The projector on the range of  $A$  is then given by

$$QQ^T = \frac{1}{5} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

Next, we see that

$$Q^T b = \frac{1}{\sqrt{5}} (1 \ 2) \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \frac{9}{\sqrt{5}}$$

and

$$R = Q^T A = \frac{1}{\sqrt{5}} (1 \ 2) \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \frac{1}{\sqrt{5}} (5 \ 9).$$

Thus, the approximate solution should satisfy the relation

$$Rx = Q^T b,$$

which implies that  $5x_1 + 9x_2 = 9$ . Note that the approximate solution is not unique (instead of the  $QR$ -decomposition). In particular, the  $QR$ -decomposition of our matrix  $A$  has the form

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \frac{1}{\sqrt{5}} (5 \ 9).$$

□

**3.E.13.** Minimise  $\|Ax - b\|$  for

$$A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Write down also the  $QR$ -decomposition of the matrix  $A$ .

**Solution.** The normalised first column of the matrix  $A$  is  $u_1 = \frac{1}{\sqrt{6}}(2, -1, -1)^T$ . From the second column, subtract



its component in the direction  $u_1$  and set  $u = (-1, 2, -1)^T$ . Then we see that  $\langle u, u_1 \rangle = -\frac{3}{\sqrt{6}}$ . Thus it follows that

$$u - \langle u, u_1 \rangle u_1 = \frac{1}{2} \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}.$$

Hence, we have created an orthogonal vector which we may normalise to obtain the vector

$$u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}.$$

The third column of the matrix  $A$  is already linearly dependent and thus the orthogonal matrix  $Q$  has the form

$$Q = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & 0 \\ -1 & \sqrt{3} \\ -1 & -\sqrt{3} \end{pmatrix}.$$

Consequently, we obtain

$$\begin{aligned} R = Q^T A &= \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \\ &= \frac{1}{\sqrt{6}} \begin{pmatrix} 6 & -3 & -3 \\ 0 & 3\sqrt{3} & -3\sqrt{3} \end{pmatrix} \end{aligned}$$

and

$$Q^T b = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

The solution of the equation  $Rx = Q^T b$  is  $x = y = z$ . Thus, multiples of the vector  $(1, 1, 1)^T$  minimize  $\|Ax - b\|$ . Note that the mapping given by the matrix  $A$ , is a projection on the plane with normal vector  $(1, 1, 1)^T$ .  $\square$



*When dealing with an overdetermined system of linear equations (more equations than unknowns), the system may not have an exact solution. In such cases, the pseudo-inverse provides the best approximate solution in the least squares sense. The pseudo-inverse is also used in linear regression for finding the optimal parameters.*

*Next we will analyze how to solve problems with linear regression. Such tasks are about finding the best approximation of some functional dependence, using a linear function.*

*In particular, given a functional dependence for some points, that is,*

$$f(a_1^1, \dots, a_n^1) = y_1, \dots, f(a_1^k, a_2^k, \dots, a_n^k) = y_k,$$

*with  $k > n$ , we wish to find the “best possible” approximation of this dependency using a linear function. Observe that since  $k > n$  one has more equations than unknowns.*

*Hence we want to express the value of the property as a linear function  $f(x_1, \dots, x_n) = b_1 x_1 + b_2 x_2 + \dots + b_n x_n + c$ .*

We choose to define “best possible” by the minimisation of

$$\sum_{i=1}^k \left( y_i - \sum_{j=1}^n (b_j x_j + c) \right)^2$$

with regard to the real constants  $b_1, \dots, b_n, c$ . The goal is to find such a linear combination of the columns of the matrix  $A = (a_j^i)$  (with coefficients  $b_1, \dots, b_n$ ), that is closest to the vector  $(y_1, \dots, y_k)^T$  in  $\mathbb{R}^k$ .

This means that the whole procedure is about specifying an orthogonal projection of the vector  $(y_1, \dots, y_k)^T$  onto the subspace generated by the columns of  $A$ . Using the theorem presented in 3.5.7, it follows that this projection should be described by the vector  $(b_1, \dots, b_n)^T = A^\dagger (y_1, \dots, b_n)$ , where  $A^\dagger$  is the pseudo-inverse of  $A$ .

**3.E.14. Linear regression.** Using the least squares method, solve the linear system given below:

$$\begin{cases} 2x + y + 2z = 1, \\ x + y + 3z = 2, \\ 2x + y + z = 0, \\ x + z = -1. \end{cases}$$

**Solution.** The system has no solution, since its matrix has rank 3, and the extended matrix has rank 4. According to the theorem in 3.5.7, the best approximation of the vector  $b = (1, 2, 0, -1)^T$  can be obtained by the vector  $A^\dagger b$ . In particular,  $AA^\dagger b$  is then the best approximation – the perpendicular projection of the vector  $b$  onto the column space  $C(A)$  of  $A$ . Because the columns of the matrix  $A$  are linearly independent, its pseudo-inverse is given by the relation  $(A^T A)^{-1} A^T$ . We compute

$$\begin{aligned} A^\dagger &= \left( \begin{pmatrix} 2 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix} \\ &= \left( \begin{pmatrix} 10 & 5 & 10 \\ 5 & 3 & 6 \\ 10 & 6 & 15 \end{pmatrix} \right)^{-1} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 3/5 & -1 & 0 \\ -1 & 10/3 & -2/3 \\ 0 & -2/3 & 1/3 \end{pmatrix} \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1/5 & -2/5 & 1/5 & 3/5 \\ 0 & 1/3 & 2/3 & -5/3 \\ 0 & 1/3 & -1/3 & 1/3 \end{pmatrix}, \end{aligned}$$

and now we can compute the desired  $x$ , which is given by

$$x = A^\dagger b = (-6/5, 7/3, 1/3)^T.$$

Consequently, the best possible approximation to the column of the right side, is the vector  $(3/5, 32/15, 4/15, -13/15)^T$ .  $\square$

F. Additional exercises for the whole chapter

As usual, we proceed with supplementary exercises related to the notions that we have discussed so far.

A) MATERIAL ON LINEAR PROGRAMMING

**3.F.1. Manufacturing bolts and nuts.** A company manufactures bolts and nuts. Moulding a box of bolts takes one minute, while a box of nuts is moulded for 2 minutes. Preparing the box itself takes one minute for bolts, 4 minutes for nuts. The company has at its disposal two hours for moulding and three hours for box preparation. Demand says that it is necessary to manufacture at least 90 boxes of bolts more than boxes of nuts. Due to technical reasons it is not possible to manufacture more than 110 boxes of bolts. The profit from one box of bolts is \$4 and the profit from one box of nuts is \$6. The company has no trouble with selling. Write down the corresponding LP problem and present its standard form. Deduce graphically how many boxes of nuts and bolts should be manufactured in order to have maximal profit.

**Solution.** For convenience, let us put the given data into a table:

	Bolts 1 box	Nuts 1 box	Capacity
Mould	1 min./box	2 min./box	160 min.
Box	1 min./box	4 min./box	180 min.
Profit	\$4/box	\$6/box	

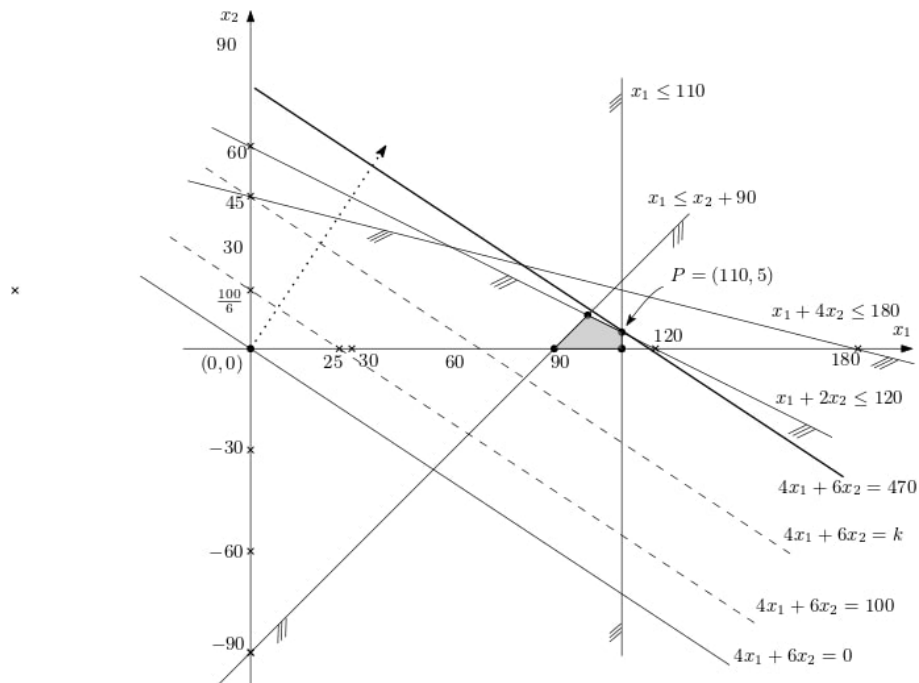
Denote by  $x_1$  the number of manufactured boxes of bolts and by  $x_2$  the number of manufactured boxes of nuts. From the restriction on moulding time and from the restriction on the box preparation we obtain the following restrictive conditions:

$$x_1 + 2x_2 \leq 120, \quad x_1 + 4x_2 \leq 180, \quad x_1 \geq x_2 + 90, \quad x_1 \leq 110.$$

The standard form reads as the maximization of the profit function  $h(x_1, x_2) = 4x_1 + 6x_2$  subject to the conditions

$$x_1 + 2x_2 \leq 120, \quad x_1 + 4x_2 \leq 180, \quad x_2 - x_1 \leq -90, \quad x_1 \leq 110,$$

with  $x_1 \geq 0, x_2 \geq 0$ . The feasible region is the grey region in the figure below, where we have also included the objective lines  $4x_1 + 6x_2 = k$ . From this figure we deduce that the point  $P = (110, 5)$  maximizes  $h$ , and the maximum possible income is thus given by  $4 \cdot 110 + 6 \cdot 5 = \$470$ .



□

**3.F.2. Investments and profits.** An insurance company has a capital of 100000€ which aims to invest in two different ways: the investment of type X and the investment of type Y. These types of investments give an annual income of 10% and of 15%, respectively. However, there is a limitation by the government, which requires that at least 25% of the capitals, should be invested in the X-type investment. On the other hand, the policy of the company requires the ratio between the capital used

for the Y-type investment and the capital used for the X-type of investment, not be greater than 1.5. How the company should invest its capital? Articulate the problem as a LP problem, and find the solution via the simplex method. Verify your solution using Sage.

**Solution.** Let  $x_1, x_2$  be the decision variables corresponding to the capitals (in euro) that the company will use for the X-type investment, and Y-type investment, respectively. The objective function is given by

$$h = 0.1x_1 + 0.15x_2 = \frac{1}{10}x_1 + \frac{3}{20}x_2 = \frac{2}{20}x_1 + \frac{3}{20}x_2.$$

The corresponding LP problem is the maximization of  $h$  under the constraints

$$x_1 + x_2 \leq 100000, \quad \frac{x_1}{x_1 + x_2} \geq 0.25, \quad \frac{x_2}{x_1} \leq 1.5, \quad x_1 \geq 0, \quad x_2 \geq 0,$$

which we can equivalently write as

$$x_1 + x_2 \leq 100000, \quad -\frac{3}{4}x_1 + \frac{1}{4}x_2 \leq 0, \quad -\frac{3}{2}x_1 + x_2 \leq 0, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

Let us introduce the slack variables  $y_1, y_2, y_3$ , with  $y_i \geq 0$  for any  $i = 1, 2, 3$ . The application of the simplex method yields four tableaux, which we present as follows:

$-h$	$x_1$	$x_2$	$y_1$	$y_2$	$y_3$	0		$-h$	$x_1$	$x_2$	$y_1$	$y_2$	$y_3$	0	
$y_1$	1	1	1	0	0	100000	$x_2 \xrightarrow{\text{enters}}$ $y_2 \xrightarrow{\text{leaves}}$	$y_1$	4	0	1	-4	0	100000	$x_1 \xrightarrow{\text{enters}}$ $y_3 \xrightarrow{\text{leaves}}$
$y_2$	-3/4	1/4	0	1	0	0		$x_2$	-3	1	0	4	0	0	
$y_3$	-3/2	1	0	0	1	0		$y_3$	3/2	0	0	-4	1	0	
$-h$	$x_1$	$x_2$	$y_1$	$y_2$	$y_3$	0		$-h$	$x_1$	$x_2$	$y_1$	$y_2$	$y_3$	13000	
$y_1$	0	0	1	20/3	-8/3	100000	$y_2 \xrightarrow{\text{enters}}$ $y_1 \xrightarrow{\text{leaves}}$	$y_2$	0	0	3/20	1	-2/5	15000	
$x_2$	0	1	0	-4	2	0		$x_2$	0	1	3/5	0	2/5	60000	
$x_1$	1	0	0	-8/3	2/3	0		$x_1$	1	0	2/5	0	-2/5	40000	

Since in the last tableau the row of  $-h$  has no negative entry, we have arrived to the optimal solution which reads as follows: ( $x_1 = 40000, x_2 = 60000$ ) with  $y_2 = 15000$ . The maximum value of  $h$  is given by 13000, hence the investments will bring to the company a maximum profit of 13000€. To verify the solution in Sage, you can use the block

```
p = MixedIntegerLinearProgram()
v = p.new_variable(real=True, nonnegative=True)
x1, x2 = v["x1"], v["x2"]
p.set_objective((2/20)*x1 + (3/20)*x2)
p.add_constraint(x1 + x2 <= 100000)
p.add_constraint((-3/4)*x1 + (1/4)*x2 <= 0)
p.add_constraint((-3/2)*x1 + x2 <= 0)
k = p.solve()
x1, x2 = p.get_values(x1, x2)
print("Answer =", round(k, 2))
print("(x1, x2) =", (x1, x2))
```

By executing this block we obtain the desired verification:

```
Answer = 13000.0
(x1, x2) = 40000.0 60000.0
```

□

**3.F.3.** If  $x_i \geq 0$  for all  $i = 1, 2, 3$ , minimize the expression  $-3x_1 - x_2 - 2x_3$  subject to the conditions

$$\begin{aligned} x_1 - x_2 + x_3 &\geq -4, \\ 2x_1 + x_3 &\leq 3, \\ x_1 + x_2 + 3x_3 &\leq 8. \end{aligned}$$

**Solution.** We multiply the objective function and the first inequality by  $-1$ , to get the equivalent task of maximizing  $h = 3x_1 + x_2 + 2x_3$  subject to the following conditions

$$\begin{aligned} -x_1 + x_2 - x_3 &\leq 4, \\ 2x_1 + x_3 &\leq 3, \\ x_1 + x_2 + 3x_3 &\leq 8, \end{aligned}$$

with  $x_i \geq 0$  for  $i = 1, 2, 3$ . Introducing the non-negative slacks  $x_4, x_5$  and  $x_6$ , interpret the objective function as  $3x_1 + x_2 + 2x_3 + 0 \cdot x_4 + 0 \cdot x_5 + 0 \cdot x_6$ . We now write down the first simplex tableau (the pivot element always appear circled):

$$\begin{array}{c|cccccc|c} R_0 & -3 & -1 & -2 & 0 & 0 & 0 & 0 \\ \hline R_1 & -1 & 1 & -1 & 1 & 0 & 0 & 4 \\ R_2 & \textcircled{2} & 0 & 1 & 0 & 1 & 0 & 3 \\ R_3 & 1 & 1 & 3 & 0 & 0 & 1 & 8 \end{array}$$

To eliminate the first column which is the driver, we apply the elementary row operations  $R_2 \rightarrow \hat{R}_2 := \frac{1}{2}R_2$ ,  $R_0 \rightarrow R_0 + 3\hat{R}_2$ ,  $R_1 \rightarrow R_1 + \hat{R}_2$ , and  $R_3 \rightarrow R_3 - \hat{R}_2$ . This yields the second tableau:

$$\begin{array}{c|cccccc|c} \hat{R}_0 & 0 & -1 & -1/2 & 0 & 3/2 & 0 & 9/2 \\ \hline \hat{R}_1 & 0 & \textcircled{1} & -1/2 & 1 & 1/2 & 0 & 11/2 \\ \hat{R}_2 & 1 & 0 & 1/2 & 0 & 1/2 & 0 & 3/2 \\ \hat{R}_3 & 0 & 1 & 5/2 & 0 & -1/2 & 1 & 13/23 \end{array}$$

where the new basic variables are  $x_1 = 3/2$ ,  $x_4 = 11/2$ , and  $x_6 = 13/2$ . This reflects the fact that we moved as much from the former slack  $x_5$  to the new basic variable  $x_1$  as possible. This increased the value of the objective function, which we may read in the right top corner of the tableau. We now move to the new work column which is the one containing  $-1$  in  $\hat{R}_0$ . In this column we circled the new pivot, which equals to 1 (since  $11/2 < 13/2$ ). Hence we can directly eliminate the second column by the row operations  $\hat{R}_0 \rightarrow \hat{R}_0 + \hat{R}_1$  and  $\hat{R}_3 \rightarrow \hat{R}_3 - \hat{R}_1$  (observe that  $\hat{R}_2$  already contains 0 in the work column). This gives

$$\begin{array}{c|cccccc|c} \tilde{R}_0 & 0 & 0 & -1 & 1 & 2 & 0 & 10 \\ \hline \tilde{R}_1 & 0 & 1 & -1/2 & 1 & 1/2 & 0 & 11/2 \\ \tilde{R}_2 & 1 & 0 & 1/2 & 0 & 1/2 & 0 & 3/2 \\ \tilde{R}_3 & 0 & 0 & \textcircled{3} & -1 & -1 & 1 & 1 \end{array}$$

and shifts the new basic variable from  $x_4$  to  $x_2$ . In this way we have increased the objective function. However, the first row  $\tilde{R}_0$  still contains a negative number and hence there is one more repetition. The new work column is the third one and the next pivot is the circled number 3. By applying the necessary elementary row operations we result to the final tableau

$$\begin{array}{c|cccccc|c} 0 & 0 & 0 & 2/3 & 5/3 & 1/3 & 31/3 \\ \hline 0 & 1 & 0 & 5/6 & 1/3 & 1/6 & 17/3 \\ 1 & 0 & 0 & 1/6 & 2/3 & -1/6 & 4/3 \\ 0 & 0 & 1 & -1/3 & -1/3 & 1/3 & 1/3 \end{array}$$

where the basic variables are the initial decision variables  $x_1, x_2, x_3$ . This gives the optimal solution  $x_1 = 4/3$ ,  $x_2 = 17/3$  and  $x_3 = 1/3$  with the maximal value of  $h$  being  $31/3$ .  $\square$

**3.F.4. Infinite optimal solutions.** Consider the LP problem of maximizing  $h = 3x_1 + x_2$  subject to the conditions

$$6x_1 + 2x_2 \leq 30, \quad 10x_1 + x_2 \leq 40, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

Use the simplex method to show that an optimal solution of this LP problem is not unique. Verify your answer by a figure.

**Solution.** Let us introduce slack variables  $y_1, y_2$  to bring the problem to its canonical form. This corresponds to the maximization of  $h = 3x_1 + x_2 + 0y_1 + 0y_2$  under the constraints

$$\begin{array}{rclcl} 6x_1 & + & 2x_2 & + & y_1 & & = & 30, \\ 10x_1 & + & x_2 & & & + & y_2 & = & 40, \end{array}$$

with  $x_i \geq 0$  and  $y_j \geq 0$  for  $i, j = 1, 2$ . By applying the simplex method we need two iterations to arrive to an optimal solution. Let us summarize the tableaux together with the pivot elements circled (we also indicate the corresponding elementary row operations).

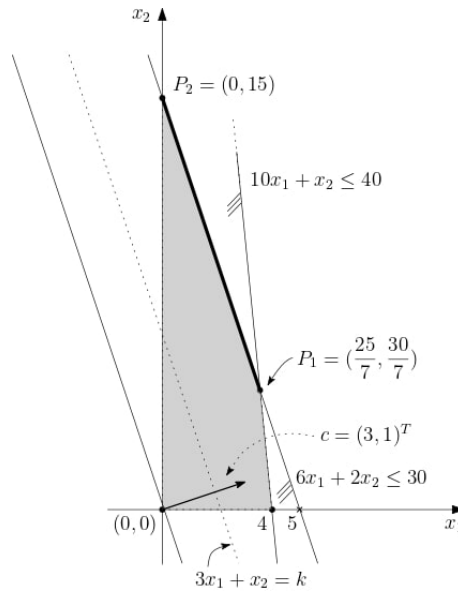
$$\begin{array}{c|cccc|c} & x_1 & x_2 & y_1 & y_2 & \\ \hline R_0 & -h & -3 & -1 & 0 & 0 & 0 \\ R_1 & y_1 & 6 & 2 & 1 & 0 & 30 \\ R_2 & y_2 & \textcircled{10} & 1 & 0 & 1 & 40 \end{array} \quad \begin{array}{l} R_2 \rightarrow \hat{R}_2 := \frac{1}{10}R_2 \\ R_0 \rightarrow \hat{R}_0 := R_0 + 3\hat{R}_2 \\ R_1 \rightarrow \hat{R}_1 := R_1 - 6\hat{R}_2 \\ \implies \end{array} \quad \begin{array}{c|cccc|c} \hat{R}_0 & -h & 0 & -7/10 & 0 & 3/10 & 12 \\ \hline \hat{R}_1 & y_1 & 0 & \textcircled{7/5} & 1 & -6/10 & 6 \\ \hat{R}_2 & x_1 & 1 & 1/10 & 0 & 1/10 & 4 \end{array}$$

$$\begin{array}{l} \hat{R}_1 \rightarrow \tilde{R}_1 := \frac{5}{7}\hat{R}_1 \\ \hat{R}_0 \rightarrow \tilde{R}_0 := \hat{R}_0 + \frac{7}{10}\tilde{R}_1 \\ \hat{R}_2 \rightarrow \tilde{R}_2 := \hat{R}_2 - \frac{1}{10}\tilde{R}_1 \\ \implies \end{array} \quad \begin{array}{c|cccc|c} & x_1 & x_2 & y_1 & y_2 & \\ \hline \tilde{R}_0 & -h & 0 & 0 & 1/2 & 0 & 15 \\ \hline \tilde{R}_1 & x_2 & 0 & 1 & 5/7 & -3/7 & 30/7 \\ \hline \tilde{R}_2 & x_1 & 1 & 0 & -1/14 & 1/7 & 25/7 \\ \hline \end{array}$$

Hence in the first iteration  $x_1$  enters the column of basic variables and  $y_2$  leaves, while in the second one  $x_2$  enters and  $y_1$  leaves. Since  $\tilde{R}_0$  does not contain a negative entry, we have arrived to an optimal solution given by  $(x_1 = \frac{25}{7}, x_2 = \frac{30}{7})$  (with  $y_1 = 0 = y_2$ ). The maximal value of  $h$  equals to 15. However, observe that in the above table the non-basic variable  $y_2$  appears with zero coefficient in  $\tilde{R}_0$ . Therefore,  $y_2$  can be brought to the basis to generate a new optimal solution with the same value of the objective function  $h$ . For this we use the entry  $1/7$  of  $\tilde{R}_2$  as a pivot and apply the following row operations  $\tilde{R}_2 \rightarrow r_2 := 7\tilde{R}_2$ ,  $\tilde{R}_1 \rightarrow \tilde{R}_1 + \frac{3}{7}r_2$ ,  $\tilde{R}_0 \rightarrow \tilde{R}_0$  (the cost row remains the same). This gives the following tableau

		$x_1$	$x_2$	$y_1$	$y_2$	
$r_0$	$-h$	0	0	1/2	0	15
$r_1$	$x_2$	3	1	1/2	0	15
$r_2$	$y_2$	7	0	-1/2	1	25

which provides the new optimal solution  $(x_1 = 0, x_2 = 15)$  (with  $y_1 = 0$  and  $y_2 = 25$ ), such that  $h(0, 15) = 15$ . Since two optimal solutions have been indicated, their convex combination will be also an optimal solution of the initial LP problem. This means that  $tP_1 + (1-t)P_2$  is also an optimal solution, for any  $t$  with  $0 \leq t \leq 1$ , where  $P_1 := (\frac{25}{7}, \frac{30}{7})$  and  $P_2 := (0, 15)$ . Thus this LP problem admits infinite optimal solutions, as one can also deduce by the help of the following figure:



□

**3.F.5. Cycling in the simplex method.** Consider the following task: maximize the functional  $h = 10x_1 - 57x_2 - 9x_3 - 24x_4$  subject to the conditions

$$\frac{1}{2}x_1 - \frac{11}{2}x_2 - \frac{5}{2}x_3 + 9x_4 \leq 0, \quad \frac{1}{2}x_1 - \frac{3}{2}x_2 - \frac{1}{2}x_3 + x_4 \leq 0, \quad x_1 + x_2 + x_3 + x_4 \leq 1, \quad x_i \geq 0, \quad (i = 1, \dots, 4).$$

Show that the simplex method applied to this LP problem *cycles*, which roughly speaking means that the same tableau occurs more than once. Can you find an optimal solution? <sup>18</sup>

**Solution.** Introduce the slack variables  $y_1, y_2, y_3$ , such that

<sup>18</sup>This example is adapted from the book of V. Chvátal (1983), *Linear Programming*, W. H. Freeman. Vasek Chvátal is a czech mathematician known for his contributions in LP, graph theory and combinatorics.

$$\begin{aligned} \frac{1}{2}x_1 - \frac{11}{2}x_2 - \frac{5}{2}x_3 + 9x_4 + y_1 &= 0, \\ \frac{1}{2}x_1 - \frac{3}{2}x_2 - \frac{1}{2}x_3 + x_4 + y_2 &= 0, \\ x_1 + x_2 + x_3 + x_4 + y_3 &= 1, \end{aligned}$$

with  $x_i \geq 0$  and  $y_j \geq 0$  for any  $i = 1, \dots, 4$  and  $j = 1, 2, 3$ . Thus the initial simplex tableau reads as follows:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$R_0$	$-h$	-10	57	9	24	0	0	0	0
$R_1$	$y_1$	$\left(\frac{1}{2}\right)$	-11/2	-5/2	9	1	0	0	0
$R_2$	$y_2$	1/2	-3/2	-1/2	1	0	1	0	0
$R_3$	$y_3$	1	1	1	1	0	0	1	1

In the driver column, which is the column of  $x_1$ , there are two choices of a pivot and we choose the circled  $1/2$  in  $R_1$ . Hence the variable  $x_1$  enters the column of basic variables and the variable  $y_1$  leaves. We replace  $R_1$  by  $\hat{R}_1 := 2R_1$  and next apply the row operations  $R_0 \rightarrow R_0 + 10\hat{R}_1$ ,  $R_2 \rightarrow R_2 - \frac{1}{2}\hat{R}_1$  and  $R_3 \rightarrow R_3 - \hat{R}_1$ . This gives the second tableau which reads as follows:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$\hat{R}_0$	$-h$	0	-53	-41	204	20	0	0	0
$\hat{R}_1$	$x_1$	1	-11	-5	18	2	0	0	0
$\hat{R}_2$	$y_2$	0	$\left(4\right)$	2	-8	-1	1	0	0
$\hat{R}_3$	$y_3$	0	12	6	-17	-2	0	1	1

Since  $-53 < -41$ , the column of  $x_2$  is the new work column and the circled 4 in  $\hat{R}_2$  is the new pivot. So  $x_2$  enters and  $y_2$  leaves. To eliminate the second column we replace  $\hat{R}_2$  by  $\tilde{R}_2 := \frac{1}{4}\hat{R}_2$  and apply the row operations  $\hat{R}_0 \rightarrow \hat{R}_0 + 53\tilde{R}_2$ ,  $\hat{R}_1 \rightarrow \hat{R}_1 + 11\tilde{R}_2$  and  $\hat{R}_3 \rightarrow \hat{R}_3 - 12\tilde{R}_2$ . This gives the third tableau:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$\tilde{R}_0$	$-h$	0	0	-29/2	98	27/4	53/4	0	0
$\tilde{R}_1$	$x_1$	1	0	$\left(\frac{1}{2}\right)$	-4	-3/4	11/4	0	0
$\tilde{R}_2$	$x_2$	0	1	1/2	-2	-1/4	1/4	0	0
$\tilde{R}_3$	$y_3$	0	0	0	7	1	-3	1	1

The column of  $x_3$  is the new driver and we choose as pivot the circled  $1/2$  in  $\tilde{R}_1$ . So  $x_3$  enters the column of basic variables and  $x_1$  leaves. To do so we replace  $\tilde{R}_1$  by  $\check{R}_1 := 2\tilde{R}_1$  and apply the row operations  $\tilde{R}_0 \rightarrow \tilde{R}_0 + \frac{29}{2}\check{R}_1$  and  $\tilde{R}_2 \rightarrow \tilde{R}_2 - \frac{1}{2}\check{R}_1$ . This gives the fourth tableau:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$\check{R}_0$	$-h$	29	0	0	-18	-15	93	0	0
$\check{R}_1$	$x_3$	2	0	1	-8	-3/2	11/2	0	0
$\check{R}_2$	$x_2$	-1	1	0	$\left(2\right)$	1/2	-5/2	0	0
$\check{R}_3$	$y_3$	0	0	0	7	1	-3	1	1

Since  $-18 < -15$ , the next work column is the column of  $x_4$  and the pivot is the circled 2 in  $\check{R}_2$ . Hence  $x_4$  enters and  $x_2$  leaves. For this, replace  $\check{R}_2$  by  $\bar{R}_2 := \frac{1}{2}\check{R}_2$  and apply the row operations  $\check{R}_0 \rightarrow \check{R}_0 + 18\bar{R}_2$ ,  $\check{R}_1 \rightarrow \check{R}_1 + 8\bar{R}_2$  and  $\check{R}_3 \rightarrow \check{R}_3 - 7\bar{R}_2$ . This gives the following table:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$\bar{R}_0$	$-h$	20	9	0	0	-21/2	141/2	0	0
$\bar{R}_1$	$x_3$	-2	4	1	0	$\left(\frac{1}{2}\right)$	-9/2	0	0
$\bar{R}_2$	$x_4$	-1/2	1/2	0	1	1/4	-5/4	0	0
$\bar{R}_3$	$y_3$	7/2	-7/2	0	0	-3/4	23/4	1	1

The new column driver is the column of  $y_1$  and we choose as pivot the circled  $1/2$  in  $\bar{R}_1$ . Thus  $y_1$  enters and  $x_3$  leaves. To do so we replace  $\bar{R}_1$  by  $r_1 := 2\bar{R}_1$  and apply the row operations  $\bar{R}_0 \rightarrow \bar{R}_0 + \frac{21}{2}r_1$ ,  $\bar{R}_2 \rightarrow \bar{R}_2 - \frac{1}{4}r_1$  and  $\bar{R}_3 \rightarrow \bar{R}_3 + \frac{3}{4}r_1$ . This yields the following table:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$r_0$	$-h$	-22	93	21	0	0	-24	0	0
$r_1$	$y_1$	-4	8	2	0	1	-9	0	0
$r_2$	$x_4$	1/2	-3/2	-1/2	1	0	1	0	0
$r_3$	$y_3$	1/2	5/2	3/2	0	0	-1	1	1

Since  $-24 < -22$  the new work column is that of  $y_2$  and the pivot is the circled 1 in  $r_2$ . Hence  $y_2$  enters the column of basic variables and  $x_4$  leaves. To eliminate the driver column we leave  $r_2$  identical and apply the row operations  $r_0 \rightarrow r_0 + 24r_2$ ,  $r_1 \rightarrow r_1 + 9r_2$  and  $r_3 \rightarrow r_3 + r_2$ . This gives us the initial simplex tableau, and hence this LP problem exhibits cycling.

A simple remedy against cycling is the so called *Bland's rule*. This says that for the entering variable we should take the first one in the present ordering of variables, which has a negative entry in the objective row. For choosing the leaving variable, if there is a tie for least ratio, take the candidate that is first in the ordering. For our problem, the ordering of the variables is  $x_1, \dots, x_4, y_1, \dots, y_3$  and an application of Bland's rule will leave everything the same up to the last pivot. This means that  $x_1$  should enter, instead of  $y_2$  and the pivot for this is the number 1/2 in  $r_2$  in the above table. Thus we replace the row  $r_2$  by  $\hat{r}_2 := 2r_2$  and apply the row operations  $r_0 \rightarrow r_0 + 22\hat{r}_2$ ,  $r_1 \rightarrow r_1 + 4\hat{r}_2$  and  $r_3 \rightarrow r_3 - \frac{1}{2}\hat{r}_2$ . This yields the following table:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$\hat{r}_0$	$-h$	0	27	-1	44	0	20	0	0
$\hat{r}_1$	$y_1$	0	-4	-2	8	1	-1	0	0
$\hat{r}_2$	$x_1$	1	-3	-1	2	0	2	0	0
$\hat{r}_3$	$y_3$	0	4	2	-1	0	-2	1	1

The next iteration will increase the value of the objective function and in particular will provide an optimal solution. Indeed,  $x_3$  enters and the new pivot is circled in the row  $\hat{r}_3$ , so  $y_3$  leaves. The final row operations are  $\hat{r}_3 \rightarrow \tilde{r}_3 := \frac{1}{2}\hat{r}_3$ ,  $\hat{r}_0 \rightarrow \hat{r}_0 + \tilde{r}_3$ ,  $\hat{r}_1 \rightarrow \hat{r}_1 + 2\tilde{r}_3$  and  $\hat{r}_2 \rightarrow \hat{r}_2 + \tilde{r}_3$  and we arrive to the following optimal table:

		$x_1$	$x_2$	$x_3$	$x_4$	$y_1$	$y_2$	$y_3$	
$\tilde{r}_0$	$-h$	0	29	0	87/2	0	19	1/2	1/2
$\tilde{r}_1$	$y_1$	0	0	0	7	1	-3	1	1
$\tilde{r}_2$	$x_1$	1	-1	0	3/2	0	1	1/2	1/2
$\tilde{r}_3$	$x_3$	0	2	1	-1/2	0	-1	1/2	1/2

Hence we have the optimal solution given by  $x_1 = 1/2 = x_3$  and  $x_2 = 0 = x_4$ , with the maximal value being  $h = 1/2$ .  $\square$

**3.F.6. Matrix games.** Imagine a game played by two players – a billionaire and fate. The billionaire would like to invest into gold, silver, diamonds or stocks of an important IT software company. Suppose that the wins and losses of such investments are well known for a period of four-five years (for simplicity, we consider only a period of four years and write them into the matrix  $A = (a_{ij})$ ):

	gold	silver	diamonds	software
2018	2%	1%	4%	3%
2019	3%	-1%	-2%	6%
2020	1%	2%	3%	-4%
2021	-2%	1%	2%	3%

The billionaire would like to invest for one year only. How should he split his investment in order to ensure the maximal win independently of development on the stock market? We assume that next year will be some (unknown) probabilistic mix of the previous four ones. In terms of our game, fate will play some stochastic vector  $(x_1, x_2, x_3, x_4)^T$  fixing the behaviour of the market (as a probabilistic mixture of the previous ones), while the billionaire will play another stochastic vector  $(y_1, y_2, y_3, y_4)^T$  describing the split of his investment. The win of the billionaire is  $\sum_{i,j=1}^4 x_i y_j a_{ij}$ .

**Solution.** The task is to find the stochastic vector  $(y_1, y_2, y_3, y_4)^T$ , which will maximize the minimum of all values

$$\sum_{i,j=1}^4 x_i y_j a_{ij}$$

for the fixed matrix  $A$  and any stochastic vector  $(x_1, x_2, x_3, x_4)^T$ . This is equivalent to the problem of maximizing  $z_1 + z_2 + z_3 + z_4$  under the condition  $A^T z \leq (1, \dots, 1)^T$ ,  $z \geq 0$  (and the requested stochastic vector  $y$  is then obtained by



normalizing the vector  $z$ , the requested optimal value is the inverse of the optimal value obtained).<sup>19</sup> Thus, one has to solve a LP problem and the first step is to introduce the slack variables  $w_1, w_2, w_3, w_4$ , and transform the problem to the standard form:

$$\max \{z_1 + z_2 + z_3 + z_4 \mid (A^T \mid E_4)(z, w) = (1, 1, 1, 1)^T\}.$$

The initial tableau of the simplex method is the following one

-1	-1	-1	-1	0	0	0	0	0	0
2	3	1	-2	1	0	0	0	0	1
1	-1	2	1	0	1	0	0	0	1
4	-2	3	2	0	0	1	0	0	1
3	6	-4	3	0	0	0	1	0	1

and after four iterations we arrive to the final table, given by

$\frac{188}{89}$	0	0	0	$\frac{25}{89}$	0	$\frac{44}{89}$	$\frac{17}{89}$	$\frac{86}{89}$
$\frac{114}{89}$	0	1	0	$\frac{18}{89}$	0	$\frac{21}{89}$	$-\frac{2}{89}$	$\frac{37}{89}$
$-\frac{146}{89}$	0	0	0	$-\frac{9}{89}$	1	$-\frac{55}{89}$	$\frac{1}{89}$	$\frac{26}{89}$
$-\frac{85}{89}$	0	0	1	$-\frac{10}{89}$	0	$\frac{18}{89}$	$\frac{11}{89}$	$\frac{19}{89}$
$\frac{78}{89}$	1	0	0	$\frac{17}{89}$	0	$\frac{5}{89}$	$\frac{8}{89}$	$\frac{30}{89}$

We can read off the optimal solution:  $z_2 = \frac{30}{89}, z_3 = \frac{37}{89}, z_4 = \frac{19}{89}, z_1 = 0$ . The optimal value (upper right corner) is  $z_1 + z_2 + z_3 + z_4 = \frac{86}{89}$ . After rescaling to a stochastic vector (multiplying with  $\frac{89}{86}$ ) we get the solution of the original problem:  $y_1 = 0, y_2 = \frac{30}{86}, y_3 = \frac{37}{86}, y_4 = \frac{19}{86}$ . with the optimal value  $\frac{89}{86}$ . □

### B) MATERIAL ON DIFFERENCE EQUATIONS

**3.F.7.** Find a real basis of solutions for the difference equation  $x_{n+4} = x_{n+3} - x_{n+2} + x_{n+1} - x_n$ .

**Solution.** The space of the solutions is a four-dimensional vector space whose generators can be obtained from the roots of the characteristic polynomial of the given equation. The characteristic equation has the form  $r^4 - r^3 + r^2 - r + 1 = 0$ . This is a so-called *reciprocal equation*, which means that the coefficients at the  $(n - k)$ -th and  $k$ -th power of  $r$  for  $k = 1, \dots, n$ , are equal. After dividing the characteristic equation by  $r^2$  (zero cannot be a root) and setting  $u := r + \frac{1}{r}$  (note that  $r^2 + \frac{1}{r^2} = u^2 - 2$ ) we obtain

$$r^2 - r + 1 - \frac{1}{r} + \frac{1}{r^2} = u^2 - u - 1 = 0.$$

This gives the indeterminates  $u_{1,2} = \frac{1 \pm \sqrt{5}}{2}$  and by the equation  $r^2 - ur + 1 = 0$  we determine the roots

$$r_{1,2,3,4} = \frac{1 \pm \sqrt{5} \pm \sqrt{-10 \pm 2\sqrt{5}}}{4}.$$

In fact, since  $r^5 + 1 = (r + 1)(r^4 - r^3 + r^2 - r + 1)$ , the roots of the characteristic equation could have been “guessed” right away. Thus, the roots of the characteristic polynomial are also the roots of the polynomial  $r^5 + 1$ , which are exactly the fifth roots of the  $-1$ . By this we obtain that the solutions of the characteristic polynomial are the numbers

$$r_{1,2} = \cos\left(\frac{\pi}{5}\right) \pm i \sin\left(\frac{\pi}{5}\right) \quad \text{and} \quad r_{3,4} = \cos\left(\frac{3\pi}{5}\right) \pm i \sin\left(\frac{3\pi}{5}\right).$$

Now, by the description in 3.2.5 it follows that a real basis of the space of the solutions of the given difference equation, is given by the sequences

$$\cos\left(\frac{n\pi}{5}\right), \sin\left(\frac{n\pi}{5}\right), \cos\left(\frac{3n\pi}{5}\right), \sin\left(\frac{3n\pi}{5}\right).$$

Note that these are sines and cosines of the arguments of the corresponding powers of the roots of the characteristic polynomial. □

<sup>19</sup>The observation comes from the proof of the von Neumann Minimax theorem, 1928. The theorem claims that any probabilistic extension of a matrix game enjoys an equilibrium state.

**3.F.8. A simplified model for the behaviour of gross domestic product.** Non homogeneous difference equations of second order are known in the science of *macroeconomics*. Consider for example the difference equation

$$x_{n+2} - A(1 + B)x_{n+1} + ABx_n = 1,$$

where  $x_k$  is the gross domestic product at the year  $k$ . The constant  $A$  is the *consumption tendency*, which is a macro economical factor that gives the fraction of money that the people spend (from what they have at their disposal). The constant  $B$  describes the dependence of the measure of investment of the private sector on the consumption tendency. Moreover, we assume that the size of the domestic product is normalised such that the right-hand side of the equation is 1. Compute the values  $x_n$  for  $A = 3/4$ ,  $B = 1/3$ ,  $x_0 = 1 = x_1$ .

**Solution.** For  $A = 3/4$  and  $B = 1/3$ , it is easy to check that the constant function  $x_n = 4$  is a particular solution of the initial (non-homogeneous) equation. We now look for solutions of the corresponding homogeneous equation. The general characteristic equation has the form

$$r^2 - A(1 + B)r + AB = 0,$$

which reduces to  $r^2 - r + 1/4 = 0$  for the particular values of  $A$  and  $B$ . Hence  $1/2$  is a double root. In this case, as we have seen before (cf. 3.B.3), the solutions of the homogeneous equation are given by  $a(1/2)^n + bn(1/2)^n$ , for some  $a, b \in \mathbb{R}$ . Hence, according to the theoretical description in 3.2.6, the solutions of the given difference equation (for  $A = 3/4$  and  $B = 1/3$ ), are expressed by

$$4 + a(1/2)^n + bn(1/2)^n.$$

Using the initial conditions  $x_0 = x_1 = 1$  we also obtain  $a = b = -3$ , and we are able to present the solution explicitly:

$$x_n = 4 - 3 \left(\frac{1}{2}\right)^n - 3n \left(\frac{1}{2}\right)^n.$$

Let us verify this result in Sage, by the method that we learned in the main part (cf. 3.B.11)

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n+2) - a(n+1) + (1/4)*a(n) - 1
initial = {a(0):1, a(1):1}
rsolve(f, a(n), initial)
```

Sage's output is the expression  $4 + (-3 * n - 3)/2 * n$ . □

**3.F.9.** Find the solution of the recurrence relation  $x_{n+2} - 6x_{n+1} + 5x_n = ne^n$ .

**Solution.** Solving first the homogeneous part we obtain:

$$x_n^{(h)} = c_1 \cdot (1)^n + c_2 \cdot 5^n.$$

To find the particular solution we can use the method of variation of the constant, described in ?? For this, we should compute the Wronski determinant:

$$W_{j+1} = \det \begin{pmatrix} 1^{j+1} & 5^{j+1} \\ 1^{j+2} & 5^{j+2} \end{pmatrix} = 4 \cdot 5^{j+1}.$$

Thus,

$$x_n = c_1 + c_2 \cdot 5^n - \frac{1}{4} \sum_{j=0}^{n-1} j e^j + \left( \sum_{j=0}^{n-1} \frac{j e^j}{4 \cdot 5^{j+1}} \right) 5^n,$$

with  $c_1, c_2 \in \mathbb{R}$ . □

**3.F.10.** Find the solution of the recurrence relation  $-x_{n+3} = 3x_{n+2} + 3x_{n+1} + x_n$ , with initial conditions  $x_1 = x_2 = x_3 = 1$ . ○

C) MATERIAL ON MODELS OF GROWTH AND ITERATED PROCESSES

*In this paragraph, the initial set of exercises addresses population growth and the Leslie model. However, it is crucial to first recognize the role that "primitive matrices" have in the theory of linear iterative models. Recall that a matrix  $A$  is called primitive if there exists a positive integer  $k$  such that  $A^k$  has only positive entries. These matrices are discussed in detail in Section 3.3.3, where further information can be found.*

**3.F.11.** Which of the matrices given below are primitive?

$$A = \begin{pmatrix} 0 & 1/7 \\ 1 & 6/7 \end{pmatrix}, \quad B = \begin{pmatrix} 1/2 & 0 & 1/3 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 1/6 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 & 0 \\ 1/4 & 0 & 1/2 \\ 3/4 & 0 & 1/2 \end{pmatrix}, \quad D = \begin{pmatrix} 1/3 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/6 & 1/6 & 1/3 \\ 1/6 & 0 & 5/6 & 2/3 \end{pmatrix}.$$

**Solution.** We see that

$$A^2 = \begin{pmatrix} 1/7 & 6/49 \\ 6/7 & 43/49 \end{pmatrix}, \quad C^3 = \begin{pmatrix} 3/8 & 1/4 & 1/4 \\ 1/4 & 3/8 & 1/4 \\ 3/8 & 3/8 & 1/2 \end{pmatrix}.$$

So the matrices  $A$  and  $C$  are primitive, since  $A^2$  and  $C^3$  are positive matrices. For  $n \in \mathbb{N}$ , the middle column of the matrix  $B^n$  is always the vector  $(0, 1, 0)^T$ , which contains the entry 0. Hence the matrix  $B$  cannot be primitive. Finally, the product

$$\begin{pmatrix} 1/3 & 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/6 & 1/6 & 1/3 \\ 1/6 & 0 & 5/6 & 2/3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ a/6 + b/3 \\ 5a/6 + 2b/3 \end{pmatrix},$$

with  $a, b \in \mathbb{R}$ , implies that the matrix  $D^2$  has in the right upper corner a zero two-dimensional (square) sub-matrix. By induction, the same property is shared by the matrices  $D^3 = D \cdot D^2$ ,  $D^4 = D \cdot D^3$ ,  $\dots$ ,  $D^n = D \cdot D^{n-1}$ , and so on. Consequently, the matrix  $D$  is not primitive.  $\square$

**3.F.12. Rabbits and their population growth.** We are again interested in the population of rabbits on a meadow, as in the problem 3.B.1. However, we now assume that the rabbits are dying after reaching the ninth year of age.<sup>20</sup> Show that according to this model the population grows approximately with the geometric sequence  $1.608^t$ .

**Solution.** Denote by  $x_1(t), x_2(t), \dots, x_9(t)$  the number of rabbits according to their age (in years), at time  $t$ . Then, the number of rabbits in individual categories are described after one year by the formula

$$x_1(t+1) = x_2(t) + x_3(t) + \dots + x_9(t), \quad x_i(t+1) = x_{i-1}(t)$$

for  $i = 2, 3, \dots, 10$ . Equivalently, we may write

$$\begin{pmatrix} x_1(t+1) \\ x_2(t+1) \\ x_3(t+1) \\ x_4(t+1) \\ x_5(t+1) \\ x_6(t+1) \\ x_7(t+1) \\ x_8(t+1) \\ x_9(t+1) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \\ x_6(t) \\ x_7(t) \\ x_8(t) \\ x_9(t) \end{pmatrix}.$$

The characteristic polynomial of the given matrix is  $\lambda^9 - \lambda^7 - \lambda^6 - \lambda^5 - \lambda^4 - \lambda^3 - \lambda^2 - \lambda - 1$ . To obtain a verification in Sage we can apply the command `A.eigenvalues()`, as usual:

```
A = matrix(QQ, [[0, 1, 1, 1, 1, 1, 1, 1, 1],
                [1, 0, 0, 0, 0, 0, 0, 0, 0],
                [0, 1, 0, 0, 0, 0, 0, 0, 0],
                [0, 0, 1, 0, 0, 0, 0, 0, 0],
                [0, 0, 0, 1, 0, 0, 0, 0, 0],
                [0, 0, 0, 0, 1, 0, 0, 0, 0],
                [0, 0, 0, 0, 0, 1, 0, 0, 0],
                [0, 0, 0, 0, 0, 0, 1, 0, 0],
                [0, 0, 0, 0, 0, 0, 0, 1, 0]])
A.eigenvalues()
```

The answer shows that  $\lambda_1 \approx 1.608$  is indeed the only positive real eigenvalue, a property that every Leslie matrix satisfies (there are also eight complex eigenvalues). In fact we can estimate this root of the characteristic polynomial very well (think why this must be smaller than  $(\sqrt{5} + 1)/2$ ).

Now, the normalized eigenvector corresponding to  $\lambda_1$  has the form (the coordinates of  $X_1$  sum to 100)

<sup>20</sup>In the original model the rabbits were immortal. Note that domesticated rabbits can live between eight to twelve years.

$$X_1 \approx (38.36, 23.85, 14.83, 9.22, 5.73, 3.56, 2.21, 1.37, 0.85)^T.$$

Consequently, according to this model the population grows approximately with the geometric sequence  $1.608^t$ .  $\square$

**3.F.13.** Consider the following Leslie model in which a farmer breeds sheep. The birth-rate of sheep depends only on their age and on average is 2 lambs per sheep between one and two years of age, 5 lambs per sheep between two and three years of age and 2 lambs per sheep between three and four years of age. Younger sheep do not deliver any lambs. Every year, half of the sheep die, uniformly distributed among all age groups. Every sheep older than four years is sent to the butchery. The farmer would like to sell (living) lambs younger than one year for their skin. What proportion of the lambs can be sold every year to ensure that the size of the herd remains the same? In what ratio will the sheep then be distributed among individual age categories?

**Solution.** The matrix of the model, without an action of the farmer, can be expressed as

$$L = \begin{pmatrix} 0 & 2 & 5 & 2 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

The farmer can influence how many sheep younger than one year stay in his herd to the next year, that is, he can influence the element  $l_{12}$  of the matrix  $L$ . Thus finally we are dealing with the model encoded by the matrix

$$L = \begin{pmatrix} 0 & 2 & 5 & 2 \\ a & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

We are looking for this  $a$  such that the matrix has the eigenvalue 1 (we know that it has only one real positive eigenvalue). The characteristic polynomial of this matrix is  $\lambda^4 - 2a\lambda^2 - \frac{5}{2}\lambda - \frac{1}{2}$ . If we require 1 as a root of this polynomial, then we get the solution  $a = \frac{1}{5}$ . Thus, the farmer can sell  $\frac{1}{2} - \frac{1}{5} = \frac{3}{10}$  of lambs that are born that year. The corresponding eigenvector for the eigenvalue 1 of the given matrix is  $(20, 4, 2, 1)^T$ , and in these ratios the population stabilises.  $\square$

**3.F.14.** Consider the Leslie population growth model for the population of rats, divided into three groups according to age: younger than one year, between one year and two years, and between two years and three years. Assume that there exist no rats older than three years. The average birth-rate of one rat in individual age categories is the following: in the first group it is zero, in the second and in the third it is 2 rats. The mortality in the second group is zero, that is, the rats that survive their first year die after three years of life. Determine the mortality in the first group, if you know that the population stagnates (the total number of rats does not change).  $\circ$

**3.F.15. Model of evolution of a whale population.** For the evolution of a population (of whales), females are important. The important factor is not age but fertility. From this point of view, one can divide the females into newborns (juvenile), that is, females who are yet fertile; young fertile females; adult females with the highest fertility, and postclimacterial females who are no longer fertile, but are still important with respect to taking care of newborns and food gathering.

We model the evolution of such a population by time. For a time unit, we choose the time it takes to reach adulthood. A newborn female who survives this interval becomes fertile. The evolution of a young female to full fertility and to post-climacterial state depends mainly on the environment. That is, the transition to the next category is a random event. Analogously, the death of an individual is also a random event. Note that a young fertile female has less children per unit interval, than an adult female. Let us not try to formalise these statements.

Denote by  $x_1(t)$ ,  $x_2(t)$ ,  $x_3(t)$ ,  $x_4(t)$  the number of juvenile, young, adult and postclimacterial females in time  $t$  respectively. The amount can be expressed as a number of individuals, but also as a number of individuals relative per unit area (population density), or as a total biomass. Denote further by  $p_1$  the probability that a juvenile female survives the unit time interval and becomes fertile. Let also  $p_2$  and  $p_3$  be the respective probabilities that a young female becomes adult and that an adult female becomes old. Another random event is the death (positively formulated: survival) of females who do not move to the next category – we denote these probabilities respectively as  $q_2$ ,  $q_3$  and  $q_4$  for young, adult and old females. Of course, Each of the numbers  $p_1, p_2, p_3, q_2, q_3, q_4$  is a probability from the interval  $[0,1]$ .

Now, a young female can survive, reach adulthood or die. These events are mutually exclusive, together they form a sure certain event and cannot be excluded. Thus,  $p_2 + q_2 < 1$ . For similar reasons  $p_3 + q_3 < 1$ . Finally, we denote by  $f_2$  and  $f_3$  the average number of daughters of a young and adult female, respectively. These parameters satisfy  $0 < f_2 < f_3$ . The expected number of newborn females in the next time interval is the sum of the daughters of young and of the adult females, that is

$$x_1(t+1) = f_2x_2(t) + f_3x_3(t).$$

Let us temporarily denote by:

- $x_{2,1}(t+1)$  the number of young females in time  $t+1$ , who were juvenile in the previous time interval.
- $x_{2,2}(t+1)$  the number of young females, who were already fertile in time  $t$ , survived that time interval, but did not move into the adulthood.

The probability  $p_1$  that a juvenile female survives the interval can be expressed by classical probability, that is, by the ratio  $x_{2,1}(t+1)/x_1(t)$ . Similarly, the probability  $q_2$  can be expressed as the ratio  $x_{2,2}(t+1)/x_2(t)$  and we finally have the relation

$$x_2(t+1) = x_{2,1}(t+1) + x_{2,2}(t+1) = p_1x_1(t) + q_2x_2(t).$$

In the same way we deduce the expected number of fully fertile females:

$$x_3(t+1) = p_2x_2(t) + q_3x_3(t),$$

and finally the expected number of postclimacterical females is

$$x_4(t+1) = p_3x_3(t) + q_4x_4(t).$$

Set now

$$A := \begin{pmatrix} 0 & f_2 & f_3 & 0 \\ p_1 & q_2 & 0 & 0 \\ 0 & p_2 & q_3 & 0 \\ 0 & 0 & p_3 & q_4 \end{pmatrix} \quad \text{and} \quad x(t) := \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{pmatrix}.$$

Then, we may rewrite the previous recurrent formulas in matrix form as  $x(t+1) = Ax(t)$ . Using this matrix difference equation it is possible to compute the expected number of females in individual categories, assuming that the distribution of the population at some initial time is known. Let us focus for example on the population of orca whales, where we may assume the following parameters:

$$\begin{aligned} p_1 &= 0.9775, & q_2 &= 0.9111, & f_2 &= 0.0043, \\ p_2 &= 0.0736, & q_3 &= 0.9534, & f_3 &= 0.1132. \\ p_3 &= 0.0452, & q_4 &= 0.9804, \end{aligned}$$

In this case the time interval is one year. If we start at the time  $t=0$  with a unit measure of young females in some unoccupied area, that is, with the vector  $x(0) = (0, 1, 0, 0)^T$ , then with the help of Sage one easily computes

$$x(1) = \begin{pmatrix} 0 & 0.0043 & 0.1132 & 0 \\ 0.9775 & 0.9111 & 0 & 0 \\ 0 & 0.0736 & 0.9534 & 0 \\ 0 & 0 & 0.0452 & 0.9804 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.0043 \\ 0.9111 \\ 0.0736 \\ 0 \end{pmatrix},$$


and

$$x(2) = \begin{pmatrix} 0 & 0.0043 & 0.1132 & 0 \\ 0.9775 & 0.9111 & 0 & 0 \\ 0 & 0.0736 & 0.9534 & 0 \\ 0 & 0 & 0.0452 & 0.9804 \end{pmatrix} \begin{pmatrix} 0.0043 \\ 0.9111 \\ 0.0736 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.01224925 \\ 0.83430646 \\ 0.13722720 \\ 0.00332672 \end{pmatrix}.$$

The reader may try a computation and graphical depiction of the results for a different initial distribution of the population. The result should be an observation that the total population grows exponentially, but the ratios of the sizes of individual groups stabilise on constant values. In fact, the matrix  $A$  has the following eigenvalues

$$\lambda_1 = 1.025441326, \lambda_2 = 0.980400000, \lambda_3 = 0.834222976, \lambda_4 = 0.004835698.$$

The eigenvector associated with  $\lambda_1$  is given by  $w = (0.03697187, 0.31607121, 0.32290968, 0.32404724)^T$ , and this vector is normalized such that the sum of its components equals to the unit.

 We will now explore a variety of further exercises related to discrete Markov chains. For the reader's convenience, we have included some classic problems that lend themselves to a Markov process interpretation, such as the "ruin of a player" (see 3.F.21), the "algorithm for determining the importance of web pages" (see 3.F.20), and other. We begin with an enjoyable problem.

**3.F.16. Students into three groups.** Suppose that we can divide students into three groups, as follows: Those that are present at a lecture and pay attention, those that are present but pay no attention, and those who are in a pub instead. Now observe, lecture after lecture, how the numbers in the individual groups change.

The first step is to observe what are the probabilities that a student changes his state. Suppose that this goes as follows. For a student who pays attention: with probability 50% stays in the same state, with 40% stops paying attention and with 10% moves to the pub. For a student who pays no attention: starts paying attention with 10%, with 50% stays in the same state and with 40% moves to the pub. Finally, for a student who is in the pub we assume that there is zero probability of

returning to the lectures. How does the model evolve in time? How does the situation change if we assume at least ten percent probability that a student returns from the pub to the lecture (but is not going to pay any attention)?

**Solution.** This is obviously a (homogeneous) discrete Markov process. Its transition matrix  $T$  has the form

$$T = \begin{pmatrix} 0.5 & 0.1 & 0 \\ 0.4 & 0.5 & 0 \\ 0.1 & 0.4 & 1 \end{pmatrix},$$

where the first column is about the students paying attention, the second one about those no paying attention and the third one for those in the pub. We will prove that all the students end up in the pub. Indeed, by Sage we see that  $T$  has the following three eigenvalues:  $\lambda_1 = 1$  and the rest two are given by 0.3 and 0.4. We also see that the eigenvector associated to  $\lambda_1$  is a multiple of  $(0, 0, 1)^T$ . This vector describes the distribution of students into the three groups with the passing of time and this proves our claim (of course, such a result is clear even without any computation, as the probability of returning from the pub is zero, all students end up in the pub).

For the second task, we will show that adding 10 percent possibility for leaving the pub, will only slightly change the situation. The corresponding transition matrix is given by

$$T = \begin{pmatrix} 0.5 & 0.1 & 0 \\ 0.4 & 0.5 & 0.1 \\ 0.1 & 0.4 & 0.9 \end{pmatrix}.$$

We see by Sage (or by Mathematica), that  $\lambda_1 = 1$  is an eigenvalue of  $T$ , with corresponding eigenvector given by  $(1, 5, 21)^T$ . This means that the distribution of the students in the individual groups is encoded by the multiple of this vector such that the coordinates sum to one, that is, the vector  $(\frac{1}{27}, \frac{5}{27}, \frac{21}{27})$ . Therefore again most of the students end up in the pub.  $\square$

**3.F.17. Daily running.** Jonas goes running every evening. He has three tracks – short, middle and long. Whenever he chooses a short track, the next day he feels bad about it and chooses with equal probabilities between long and medium. Whenever he chooses a long track, the next day he chooses arbitrarily among all three. Whenever he chooses the medium track, the next he feels good about it, and again chooses with equal probabilities between medium and long. Joans claims that he has been running in this mode for a very long period. How often does he choose the short track and how often the long track? What is the probability that Jonas follows a long track, under the assumption that he did the same a week ago?

**Solution.** Clearly this task forms a (homogeneous) discrete Markov process with a three-dimensional state space:  $S = \{\text{short track, medium track, long track}\}$  (that silently one can encode as  $\{1, 2, 3\}$ ). This order of the states gives the following transition matrix

$$T = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \end{pmatrix}.$$

Let explain the computation for the second column, which corresponds to the choice of the medium track during the previous day. This means that with the probability  $1/2$ , a medium track will be chosen (the second row), and with probability  $1/2$  a long track will be chosen (the third row). Hence  $T$  follows. Now, we see that

$$T^2 = \begin{pmatrix} 1/6 & 1/6 & 1/9 \\ 5/12 & 5/12 & 4/9 \\ 5/12 & 5/12 & 4/9 \end{pmatrix}.$$

Therefore, we can use the corollary of the Perron-Frobenius theorem for Markov chains, see 3.3.4. The eigenvector corresponding to the eigenvalue 1 is the stochastic vector  $(\frac{1}{7}, \frac{3}{7}, \frac{3}{7})^T$ , where recall that the numbers  $1/7, 3/7, 3/7$  are respectively the probabilities that in a randomly chosen day, Joans choose a short, medium or long track. Suppose now that on a certain day (that is, in time  $n \in \mathbb{N}$ ), Jonas follows a long track. This corresponds to the probabilistic vector  $x^{(n)} = (0, 0, 1)^T$ . As we know, for the following day we will have

$$x^{(n+1)} = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/2 & 1/2 & 1/3 \\ 1/2 & 1/2 & 1/3 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}.$$

In particular, after seven days

$$x^{(n+7)} = T^7 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = T^6 \cdot \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}.$$

Let us present this computation in Sage:

```
T=matrix(QQbar, [[0, 0, 1/3], [1/2, 1/2, 1/3], [1/2, 1/2, 1/3]])
x=vector(QQbar, [1/3, 1/3, 1/3]); T^6*x
```

Sage prints out  $(4999/34992, 29993/69984, 29993/69984)$  and hence we deduce that

$$x^{(n+7)} \approx (4999/34992 \approx 0.142861, 29993/69984 \approx 0.428569, 29993/69984 \approx 0.428569)^T.$$

Consequently, the probability that Jonas follows a long track under the assumption that he did the same a week ago, is  $0.428569 \approx 3/7$ .  $\square$

**3.F.18. Car rental.** A car rental company has two branches – one in Prague and one in Brno. A car rented in Brno can be returned in Prague and vice versa. After operating for some time, the company has observed that roughly 80% of the cars rented in Prague and 90% of the cars rented in Brno, are finally returned in Prague. However, the strategy of the company says that in both branches, in the start of every week should exist the same number of the cars, as in the week before. How the cars should be then distributed?

**Solution.** Let us denote by  $x_B$  and  $x_P$ , the initial number of cars in Brno and in Prague, respectively. Then  $x = (x_B, x_P)^T$  is the vector describing car's distribution between the two branches. Now, according to the statement the matrix describing the (linear) system of car rental is given by

$$A = \begin{pmatrix} 0.1 & 0.2 \\ 0.9 & 0.8 \end{pmatrix}.$$

The state at the end of the week is given by  $Ax$ . However, at the end of the week in the branches the same number of cars should exist as at the beginning. This means that  $x$  must satisfy  $Ax = x$ , and so it must be an eigenvector of the matrix  $A$  associated with the eigenvalue  $\lambda_1 = 1$ . We can save time on computations by using Sage as follows:

```
A = matrix(QQ, [[0.1, 0.2], [0.9, 0.8]])
print(A.eigenvalues())
show(A.eigenvectors_right())
```

Sage's output for the eigenvector has the form

$$\left[ \left( 1, \left[ \left( 1, \frac{9}{2} \right) \right], 1 \right), \left( -\frac{1}{10}, [(1, -1)], 1 \right) \right],$$

and we deduce that the eigenvector corresponding to  $\lambda_1$  is given by  $(1, 9/2)^T$ . This implies that  $x = (1, 9/2)^T$ .

Now, the percentage distribution of the cars is given by the normalized vector associated to  $x$  and this is the vector whose entries sum to 1. This vector is given by

$$\frac{2}{11} \begin{pmatrix} 1 \\ 9/2 \end{pmatrix} = \begin{pmatrix} 0.181818 \\ 0.818182 \end{pmatrix}.$$

Therefore, the optimal distribution of cars would have approximately 18% stationed in Brno and 82% in Prague.  $\square$

**3.F.19.** Suppose that two students A and B spend every Monday morning playing a certain computer game. The person who wins pays for both of them in the evening in the restaurant. The game can also be a draw – then each pays for half the meal. The result of the previous game partially determines the next game. If a week ago student A has won, then with the probability  $3/4$  wins again and with probability  $1/4$  it is a draw. A draw is repeated with probability  $2/3$ . With probability  $1/3$  the next game is won by B. Moreover, if student B won a game, then with probability  $1/2$  he/she wins again and with probability  $1/4$ , student A wins the next game. Determine the probability that today each of them pays half of the costs, if the first game played long time ago and was won by A.

**Solution.** This is a Markov process with the following three states: "student A wins", "the game ends with a draw", "student B wins". Labelling these states in this order as  $\{1, 2, 3\}$  we arrive to the following transition matrix of the process:

$$T = \begin{pmatrix} 3/4 & 0 & 1/4 \\ 1/4 & 2/3 & 1/4 \\ 0 & 1/3 & 1/2 \end{pmatrix}.$$

We want to find the probability of the transition from the first state to the second after a large number  $n \in \mathbb{N}$  of steps (weeks). Observe that the matrix  $T$  is primitive, because

$$T^2 = \begin{pmatrix} 9/16 & 1/12 & 5/16 \\ 17/48 & 19/36 & 17/48 \\ 1/12 & 7/18 & 1/3 \end{pmatrix}.$$

Thus, it suffices to find the probabilistic eigenvector  $x_\infty$  of the matrix  $T$  associated with the eigenvalue  $\lambda_1 = 1$ . As before, giving the following block in Sage we compute the eigenvalues and eigenvectors corresponding to  $T$ :

```
T = matrix(QQbar, [[3/4, 0, 1/4], [1/4, 2/3, 1/4], [0, 1/3, 1/2]])
T.eigenvalues()
T.eigenvectors_right()
```

We deduce that the eigenvector associated to the eigenvalue  $\lambda_1 = 1$  is the vector  $(1, 3/2, 1)^T$  which we may normalize as  $x_\infty = (\frac{2}{7}, \frac{3}{7}, \frac{2}{7})^T$ . Recall now that  $x_\infty$  differs only very slightly from the probabilistic vector for large  $n$  and in particular it does not depend on the initial state. Indeed, for large  $n \in \mathbb{N}$  we compute

$$T^n \approx \begin{pmatrix} 2/7 & 2/7 & 2/7 \\ 3/7 & 3/7 & 3/7 \\ 2/7 & 2/7 & 2/7 \end{pmatrix}.$$

The desired probability is the element of this matrix on the second position in the first column (the second component of the vector  $x_\infty$ ). Hence the answer is  $3/7$ .  $\square$

**3.F.20. Algorithm for determining the importance of pages.** Internet browsers can find (almost) all pages containing a given word or phrase on the Internet. But how can a user sort the pages such that a list is sorted according to the relevance of the given pages? One of the possibilities is the following algorithm: the collection of all found pages is considered to be a system, and each of the found pages is one of its states. We describe a random walk on these pages as a Markov process.

The transition probabilities between web pages are determined by hyperlinks: each link from page A to page B establishes the probability of moving from A to B as

$$\frac{1}{\text{total number of links from page A}}.$$

If a page has no outgoing links, it is treated as a page that links to every other page. This creates a probability matrix  $M = (m_{ij})$ , where each entry  $m_{ij}$  represents the probability of moving from the  $i$ th page to the  $j$ th page.

Assume that a user randomly clicks on links, and chooses a random page when encountering a linkless page. In this case, the probability of being on the  $i$ th page at a sufficiently large time from the beginning, is given by the  $i$ th component of the unit eigenvector of the matrix  $M$ , corresponding to the unit eigenvalue. The importance of individual pages is then determined by the magnitudes of these probabilities.

This algorithm can be modified by assuming that the user eventually stops clicking on links after some time and starts again from a random page. Suppose that the user chooses a new page randomly with probability  $d$ , and with probability  $1 - d$ , continues clicking on links. In this scenario, the probability  $m_{ij}$  of transitioning from the page  $S_i$  to page  $S_j$  is given by

$$m_{ij} = \begin{cases} \frac{d}{n} + \frac{(1-d)}{\text{total number of links at the page } S_i}, & \text{if from } S_i \text{ there is a link to } S_j, \\ \frac{d}{n}, & \text{otherwise.} \end{cases}$$

Note that  $m_{ij} \neq 0$ . Now, According to the Perron-Frobenius theorem, the eigenvalue 1 is with multiplicity one and dominant, and thus the corresponding eigenvector is unique. For an illustration, consider pages A, B, C and D. The links lead from A to B and to C, from B to C and from C to A, and finally from D nowhere. Suppose that the probability that the user chooses a random new page is  $1/5$ . Then the matrix  $M$  looks as follows:

$$M = \begin{pmatrix} 1/20 & 1/20 & 17/20 & 1/4 \\ 9/20 & 1/20 & 1/20 & 1/4 \\ 9/20 & 17/20 & 1/20 & 1/4 \\ 1/20 & 1/20 & 1/20 & 1/4 \end{pmatrix}.$$

We find that the eigenvector corresponding to the eigenvalue 1 is the vector  $u = (305/53, 175/53, 315/53, 1)^T \in \mathbb{R}^4$ . Hence, the importance of the pages is given according to the order of the sizes of the corresponding components, that is,  $C > A > B > D$ .

**3.F.21. Ruining of a player.** Two players, A and B, gamble for money repeatedly in a certain game, which can result only in a victory for one of the players. The winning probability for player A in each individual game is  $p \in [0, 1/2)$ . Both players bet always the amount of 1€. Consequently, after each game player B gives 1€ to player A with probability  $p$ , and player A gives 1€ to player B with probability  $1 - p$ . Suppose that in the start of the game that A has € $x$ , B has € $y$  and that they play, as long as they both have some money. Determine the probability that the player A will lose all of his/her money.

**Solution.** This is the so-called *ruining of a player*. It is a special Markov chain (see also the exercise Sweet-toothed gambler) with many important applications. The probability in question is given by



$$\frac{1 - \left(\frac{p}{1-p}\right)^y}{1 - \left(\frac{p}{1-p}\right)^{x+y}}.$$

We can now investigate what this value is for specific choices of  $p, x, y$ . We give an example. Suppose that the player B wants the probability that the player A will lose the amount of 1 000 000€ to be at least 0.999. Assume for instance that  $p$  is given by  $p = 0.495$ . Then, for the player B we see that is sufficient to have the amount of 346€. If  $p = 0.499$ , this amount increases to 1727€. Therefore, it is possible in big casinos that “passionate” players play almost fair games.  $\square$

**3.F.22.** In a certain game you can choose one of two opponents. The probability that you beat the better one is  $1/4$ , while the probability that you beat the worse one is  $1/2$ . But the opponents cannot be distinguished, thus you do not know which one is the better one. You await a large number of games. For each of them you can choose a different opponent. Consider the following two strategies:

1. For the first game choose the opponent randomly. If you win a game, carry on with the same opponent; if you lose a game, change the opponent.
2. For the first two games, choose an opponent randomly. Then for the next two games, if you lost both the previous games, change the opponent, otherwise stay with the same opponent.

Which of the two strategies is better?

**Solution.** Both strategies define a Markov chain. For simplicity denote the worse opponent by A and the better opponent by B. In the first case the state space has two states: “game with A” and “game with B”. In this order, we obtain the probabilistic transition matrix

$$\begin{pmatrix} 1/2 & 3/4 \\ 1/2 & 1/4 \end{pmatrix}.$$

This matrix has all of its elements positive. Thus it suffices to find the probabilistic vector  $x_\infty$ , which is associated with the eigenvalue 1. We compute  $x_\infty = \left(\frac{3}{5}, \frac{2}{5}\right)^T$ . Its components correspond to the probabilities that after a long sequence of games the opponent is A, or B. Therefore, we can expect that 60 % of the games will be played against the worse of the two opponents. Because  $0.4 = \frac{2}{5} = \frac{3}{5} \cdot \frac{1}{2} + \frac{2}{5} \cdot \frac{1}{4}$ , there will be roughly 40 % against the better of the two opponents.

For the second strategy, use the states “two games in a row with A”, and “two games in a row with B”, which lead to the probabilistic transition matrix

$$\begin{pmatrix} 3/4 & 9/16 \\ 1/4 & 7/16 \end{pmatrix}.$$

In this case we compute  $x_\infty = \left(\frac{9}{13}, \frac{4}{13}\right)^T$ . Against the worse opponent one would then play (9/4)-times more frequently than against the better one. Recall that for the first strategy it is (3/2)-times more frequently, which means that second strategy is better. Note also that for the second strategy, roughly 42, 3 % of the games are winning ones, as we can see from the relation

$$0.423 \approx \frac{11}{26} = \frac{9}{13} \cdot \frac{1}{2} + \frac{4}{13} \cdot \frac{1}{4}. \quad \square$$

**3.F.23.** In a certain country there are two television channels. From a public survey it follows that in one year  $1/6$  of the viewers of the first channel move to the second one, while  $1/5$  viewers of the second channel move to the first one. Determine the time evolution of the number of viewers watching the channels, using Markov processes. Write down a matrix of the process, and find its eigenvalues and eigenvectors.  $\circ$

**3.F.24. Daily casino routine.** A female roulette player has the following strategy: she comes at a casino to play with €10. She always bets everything she has, and she always bets on black (there are 37 numbers in the roulette, 18 black, 18 red and zero). The player ends whenever she has either nothing, or when she wins €80. Consider this problem as a Markov process and write down its transition matrix.

**Solution.** In the course of the game and at its end, the player can have only one of the following amounts of money (in €): 0, 10, 20, 40, 80. If we view the situation as a Markov process, then these amounts corresponds to the states of the process, and we construct the following transition matrix

$$A = \begin{pmatrix} 1 & a & a & a & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & b & 0 & 0 & 0 \\ 0 & 0 & b & 0 & 0 \\ 0 & 0 & 0 & b & 1 \end{pmatrix},$$

with  $a = \frac{19}{37}$  and  $b = \frac{18}{37}$ , respectively.

Note that the matrix is probabilistic and singular, while the eigenvalue  $\lambda_1 = 1$  has multiplicity two (and this is the unique non-zero eigenvalue of  $A$ ), as we see from the following block in Sage:

```
a = var('a'); b = var('b')
assume(a, 'real'); assume(b, 'real')
A = matrix(SR, 5, 5, [[1, a, a, a, 0], [0, 0, 0, 0, 0], [0, b, 0, 0, 0],
[0, 0, b, 0, 0], [0, 0, 0, b, 1]])
A.eigenvalues()
```

Executing this block we obtain the list  $[0, 0, 0, 1, 1]$ . Observe that the game does *not* converge to a single vector  $x_\infty$ , but ends in one of the eigenvectors associated with the eigenvalue  $\lambda_1$ , that is, either  $(1, 0, 0, 0, 0)^T$  (the player loses it all), or  $(0, 0, 0, 0, 1)^T$  (the player wins €80). We verify this as follows (after introducing the matrix  $A$  as above)

```
A.eigenvectors_right()
```

In this case Sage output has the form

```
[(0, [(1, 0, 0, -1/a, b/a)], 3), (1, [(1, 0, 0, 0, 0), (0, 0, 0, 0, 1)], 2)]
```

Furthermore, using Sage we can easily check that the game ends after three bets, that is, the sequence  $\{A^n\}_{n=1}^\infty$ , is constant for  $n \geq 3$ . For instance, you may type

```
A^3 == A^4
A^3 == A^5
```

and for both cases Sage prints out True (and similarly for any other power  $A^n$  with  $n \geq 3$ ). Moreover, we see that  $A^3$  is given by

```
[ 1 (a*b + a)*b + a      a*b + a      a      0]
[ 0      0      0      0      0]
[ 0      0      0      0      0]
[ 0      0      0      0      0]
[ 0      0      b^3      b^2      b      1]
```

which means that

$$A^\infty := A^3 = A^n = \begin{pmatrix} 1 & a + ab + ab^2 & a + ab & a & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & b^3 & b^2 & b & 1 \end{pmatrix}.$$

We deduce that the game ends with the probability  $a + ab + ab^2 \doteq 0.885$  as a loss, and with the probability roughly 0.115 as a win of €80. This conclusion occurs by multiplying the matrix  $A^\infty$  with the initial vector  $(0, 1, 0, 0, 0)^T$ . This gives the vector  $(a + ab + ab^2, 0, 0, 0, b^3)^T$ . Notice that whether the player was female or male, makes no difference to the result.  $\square$

**3.F.25.** Consider the situation from the previous case and assume that the probability of both win and loss is  $1/2$ . Denote by  $A$  the matrix of the process. Without using any computational software determine  $A^{100}$ .  $\circ$

**3.F.26. Reliable products.** A production line is not reliable: individual products differ in quality in a significant way. A certain worker tries to improve the quality of the products and intervenes to the process. The products are distributed into classes I, II, and III, according to their quality, and a report found out the following:

- After a product of class I, the next product has the same quality in 80% of the cases and is of quality II in 10% of the cases.
  - After a product of class II, the next product is of class II in 60% of the cases and is of quality I in 20% of the cases.
  - After a product of class III, the next product is of quality III in 50% of the cases. while in 25% of the cases it is of class II.
- Compute the probability that the 18th product has quality of class I, given that the 16th product is of quality III.

**Solution.** There are at least two ways to provide an answer. First we solve the problem without using a Markov chain. Since the 16th product has quality of class III, the event in question is satisfied by the cases

- The 17th product has quality of class I and the 18th product has quality of class I,
- The 17th product has quality of class II and the 18th product has quality of class I,
- The 17th product has quality of class III and the 18th product has quality of class I,

with probabilities respectively

$$0.25 \cdot 0.8 = 0.2, \quad 0.25 \cdot 0.2 = 0.05, \quad 0.5 \cdot 0.25 = 0.125.$$

Thus the solution is  $0.375 = 0.2 + 0.05 + 0.125$ .

Let us now view the problem as a discrete Markov process. From the statement it follows that the transition matrix has the form

$$T = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix}.$$

The situation that the product is in class III is given by the probabilistic vector  $(0.0.1)^T$ . For the next product we obtain the probabilistic vector

$$\begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Finally, for the next product in order, we compute the vector

$$\begin{pmatrix} 0.375 \\ 0.3 \\ 0.325 \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.25 \\ 0.1 & 0.6 & 0.25 \\ 0.1 & 0.2 & 0.5 \end{pmatrix} \begin{pmatrix} 0.25 \\ 0.25 \\ 0.5 \end{pmatrix},$$

and the first component is the desired probability.

Observe that the first method (without using the Markov process) led to the result faster and easier. But notice also how unclear it would become if we wanted to compute, say, the 22nd or 30th product. For the second method one can in a sense restrict the computations to the relevant parts of the matrices only, instead of mindlessly multiplying the whole matrix. When using the Markov process, we have also directly obtained the probabilities that the 18th product belongs to the class II and III.  $\square$

**3.F.27.** Suppose that there are two boxes, which together contain  $n$  white and  $n$  black balls. Each box can contain up to  $n$  balls. At regular time intervals, a ball is taken from each box and moved to the other box. For this Markov process, find its probabilistic transition matrix  $T$ .

**Solution.** This problem is often used in physics as a model for blending two incompressible liquids (already introduced by D. Bernoulli during 1769), or analogously, as a model of diffusion of gases.

Let the states  $0, 1, \dots, n$  correspond to the number of white balls in the first box. This information already says how many black balls are in the first box (the remaining balls are then in the second box). If, for a certain step, the state changes from  $j \in \{1, \dots, n\}$  to  $j - 1$ , then from the first box a white ball was drawn and from the second a black ball was drawn. This happens with probability

$$\frac{j}{n} \cdot \frac{j}{n} = \frac{j^2}{n^2}.$$

The transition from state  $j \in \{0, \dots, n - 1\}$  to the state  $j + 1$  corresponds to drawing the black ball from the first box, and a white ball from the second box, with probability

$$\frac{n - j}{n} \cdot \frac{n - j}{n} = \frac{(n - j)^2}{n^2}.$$

The system stays in state  $j \in \{1, \dots, n - 1\}$ , if from both boxes balls of the same colour are drawn, which has the same probability, namely

$$\frac{j}{n} \cdot \frac{n - j}{n} + \frac{n - j}{n} \cdot \frac{j}{n} = \frac{2j(n - j)}{n^2}.$$

Notice that from the state 0 it is necessary (with probability 1) to go to the state 1 and similarly from the state  $n$  with probability one to the state  $n - 1$ . In summary, for the order of the states  $0, 1, \dots, n$ , we obtain the following  $n \times n$  transition matrix:

$$T = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ n^2 & 2 \cdot 1(n - 1) & 2^2 & \ddots & 0 & 0 & 0 \\ 0 & (n - 1)^2 & 2 \cdot 2(n - 2) & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 2 \cdot (n - 2) & (n - 1)^2 & 0 \\ 0 & 0 & 0 & \ddots & 2^2 & 2 \cdot (n - 1) & n^2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

In physics we are of course interested in the distribution of balls in boxes, after a certain time (the number of drawings). If the initial state is for instance 0, we can use this model and the powers of the matrix  $T$  to observe with what probability the number of white balls in the first box is increasing. We can confirm the expected result that the initial distribution of the balls influences their distribution, after a certain time in a very negligible way.  $\square$

D) MATERIAL ON LINEAR ALGEBRA – UNITARY SPACES, ORTHOGONALITY, JORDAN FORMS

**3.F.28. The Cauchy–Schwarz inequality.** Let  $(V, \langle \cdot, \cdot \rangle)$  be a unitary space over  $\mathbb{R}$ . Describe an alternative proof of the Cauchy inequality, in comparison with the one presented in 3.4.3.

**Solution.** We should prove the inequality  $|\langle x, y \rangle| \leq \|x\| \|y\|$ , for any two vectors  $x, y \in V$ . We may assume that  $y \neq 0$  since otherwise the statement is trivial. For some scalar  $t \in \mathbb{R}$  define the function  $\zeta(t) = \|ty - x\|^2 = \langle ty - x, ty - x \rangle$ . Then we get

$$\zeta(t) = t^2 \|y\|^2 - 2t \langle y, x \rangle + \|x\|^2.$$

Since  $(V, \langle \cdot, \cdot \rangle)$  is a real unitary space we have  $\langle y, x \rangle = \langle x, y \rangle$ , and hence  $\zeta(t) = t^2 \|y\|^2 - 2t \langle x, y \rangle + \|x\|^2$ . When  $y \neq 0$ , by setting  $a = \|y\|^2$ ,  $b = 2\langle y, x \rangle$  and  $c = \|x\|^2$  we result to a second order polynomial given by  $\zeta(t) = at^2 - bt + c$ . However,  $\|\cdot\|$  is a norm and thus  $\|ty - x\|^2 \geq 0$ , that is,  $\zeta(t) \geq 0$ . Therefore the discriminant of  $\zeta(t)$  must be non-positive:  $\Delta = b^2 - 4ac \leq 0$ , which is equivalent to  $4ac \geq b^2$ . This means  $4\|y\|^2 \|x\|^2 \geq 4\langle y, x \rangle^2$  or equivalently  $\|y\| \|x\| \geq |\langle x, y \rangle|$  and the proof is complete.  $\square$

*The next problem is about the so-called polarization identities on a unitary space. According to these formulas one can recover the inner product from the norm. The polarization identities can be used in many situations, see for example 3.F.32. (and see also Chapter 4 for a description in terms of quadratic forms).*

**3.F.29. Polarization identities.** Let  $(V, \langle \cdot, \cdot \rangle)$  be a unitary space.

(i) If  $V$  is defined over  $\mathbb{R}$  and so  $(V, \langle \cdot, \cdot \rangle)$  is a real unitary space, prove that

$$4\langle u, v \rangle = \|u + v\|^2 - \|u - v\|^2, \quad u, v \in V.$$

(ii) If  $V$  is defined over  $\mathbb{C}$  and so  $(V, \langle \cdot, \cdot \rangle)$  is a complex unitary space, prove that

$$4\langle u, v \rangle = \|u + v\|^2 - \|u - v\|^2 + i\|u + iv\|^2 - i\|u - iv\|^2, \quad u, v \in V.$$

**Solution.** (i) By assumption,  $(V, \langle \cdot, \cdot \rangle)$  is a real inner product space, hence  $\langle u, v \rangle = \langle v, u \rangle$  and by bilinearity we get

$$\|u + v\|^2 - \|u - v\|^2 = \langle u + v, u + v \rangle - \langle u - v, u - v \rangle = \|u\|^2 + 2\langle u, v \rangle + \|v\|^2 - (\|u\|^2 - 2\langle u, v \rangle + \|v\|^2) = 4\langle u, v \rangle.$$

(ii) In this case  $(V, \langle \cdot, \cdot \rangle)$  is a complex inner product space and thus  $\langle u, v \rangle \neq \langle v, u \rangle$ . We compute

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v \rangle = \|u\|^2 + \langle u, v \rangle + \langle v, u \rangle + \|v\|^2, \\ -\|u - v\|^2 &= -\langle u - v, u - v \rangle = -\|u\|^2 + \langle u, v \rangle + \langle v, u \rangle - \|v\|^2, \\ i\|u + iv\|^2 &= i\langle u + iv, u + iv \rangle = i\|u\|^2 + i\langle u, iv \rangle + i\langle iv, u \rangle + i\langle iv, iv \rangle = i\|u\|^2 + i\bar{i}\langle u, v \rangle + i^2\langle v, u \rangle + i^2\bar{i}\|v\|^2 \\ &= i\|u\|^2 + \langle u, v \rangle - \langle v, u \rangle + i\|v\|^2, \\ -i\|u - iv\|^2 &= -i\langle u - iv, u - iv \rangle = -i\|u\|^2 + \langle u, v \rangle - \langle v, u \rangle - i\|v\|^2. \end{aligned}$$

Hence adding these relations we arrive to the desired identity:  $4\langle u, v \rangle = \|u + v\|^2 - \|u - v\|^2 + i\|u + iv\|^2 - i\|u - iv\|^2$ .  $\square$

*Beyond the theory of unitary spaces the notion of norm appears in many other areas of mathematics. For instance, we couldn't imagine the so-called theory of normed vector spaces without a norm. This is a topic that we will analyze in Chapter 7, but since we are already familiar with scalar products we can present a few exercises on norms induced by scalar products. In this point an important remark is that not all the norms are of this type, see the Problem 3.F.30 which establishes a very first link with the material that we are going to treat in Chapter 7.*



**3.F.30. Parallelogram law.** Prove that the following relation makes sense on any unitary vector space  $(V, \langle \cdot, \cdot \rangle)$ :

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2), \quad (\dagger)$$

for any  $u, v \in V$ , where  $\|\cdot\|$  denotes the induced norm by  $\langle \cdot, \cdot \rangle$ . Then, show with an example that there are norms which are not induced by an inner product.

**Solution.** Many students in a high-school can solve the first task, at least for the real vector space  $\mathbb{R}^n$ . It occurs by the properties of a scalar product and the details of the proof are left to the reader. The second task is a bit more demanding. Consider for example  $\mathbb{R}^2$  and set

$$\|(x_1, x_2)\|_\infty := \max\{|x_1|, |x_2|\}$$

for any  $x = (x_1, x_2)^T \in \mathbb{R}^2$ . We can easily show that this is a norm since it satisfies the characteristic properties of a norm. For example the triangle inequality occurs as follows:

$$\|x + y\|_\infty = \max\{|x_1| + |y_1|, |x_2| + |y_2|\} \leq \max\{|x_1|, |x_2|\} + \max\{|y_1|, |y_2|\} = \|x\|_\infty + \|y\|_\infty.$$

However, this norm is not induced by an inner product on  $\mathbb{R}^2$ , since it does not satisfy the relation  $(\dagger)$  posed above. To understand this, consider for example the vectors  $x = (1, 0)^T$  and  $y = (0, 1/2)^T$ . Then we compute  $\|x + y\|_\infty^2 + \|x - y\|_\infty^2 = 2$  but  $2(\|x\|_\infty^2 + \|y\|_\infty^2) = 5/2$ .  $\square$

**3.F.31.** Let  $F : \text{Mat}_2(\mathbb{R}) \rightarrow \mathbb{R}_2[x]$  the mapping defined by

$$F \left( \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right) = (a + d) + (a + b)x + 2(c + d)x^2, \quad a, b, c, d \in \mathbb{R}.$$

After verifying that  $F$  is linear, find the dimension of the null space  $\ker(F) \subset \text{Mat}_2(\mathbb{R})$  of  $F$ . Next provide an orthonormal basis of  $\ker(F)$ , with respect to the Frobenius scalar product  $\mathcal{B}(A, B) = \text{tr}(B^T A)$  on  $\text{Mat}_2(\mathbb{R})$ .  $\circ$

**3.F.32. Linear isometries.** Let  $(V, \langle \cdot, \cdot \rangle_V)$  and  $(W, \langle \cdot, \cdot \rangle_W)$  be two (finite-dimensional) unitary spaces. For a linear mapping  $\psi : V \rightarrow W$  prove the following:

(a)  $\psi$  is a linear isometry if and only if  $\|u\|_V = \|\psi(u)\|_W$  for any  $u \in V$ , i.e.,  $\psi$  is distance preserving.

(b)  $\psi$  is a linear isometry if and only if  $\psi^* \psi = I_V$ , where  $I_V$  is the identity map on  $V$ .

**Solution.** (a) Let  $u \in V$  and assume that  $\psi$  is an isometry. This means that  $\langle u, v \rangle_V = \langle \psi(u), \psi(v) \rangle_W$  for any  $u, v \in V$  and we see that  $\|\psi(u)\|_W^2 = \langle \psi(u), \psi(u) \rangle_W = \langle u, u \rangle_V = \|u\|_V^2$ , and hence  $\|\psi(u)\|_W = \|u\|_V$  for any  $u \in V$ . Conversely, assume that  $\|u\|_V = \|\psi(u)\|_W$  for any  $u \in V$ . Then, recall by 3.F.29 that the scalar product is defined by the norm via the formula

$$4\langle u, v \rangle_V = \|u + v\|_V^2 - \|u - v\|_V^2 + i\|u + iv\|_V^2 - i\|u - iv\|_V^2,$$

for any  $u, v \in V$ , and similarly for  $\langle \cdot, \cdot \rangle_W$ . Because  $\psi$  preserves norms, a direct calculation via the above formula certifies that it will preserve also the inner products, and so  $\psi$  is an isometry.

(b) Suppose that  $\psi$  satisfies  $\psi^* \psi = I_V$ . Then, for any  $u, v \in V$  we have

$$\|u - v\|_V^2 = \langle u - v, u - v \rangle_V = \langle u - v, (\psi^* \psi)(u - v) \rangle_V = \langle \psi(u - v), \psi(u - v) \rangle_W = \|\psi(u) - \psi(v)\|_W^2.$$

Thus  $\|u - v\|_V = \|\psi(u) - \psi(v)\|_W$  which shows that  $\psi$  is an isometry. Conversely, if  $\psi : V \rightarrow W$  is an isometry, then  $\langle (\psi^* \psi)(u), v \rangle_V = \langle \psi(u), \psi(v) \rangle_W = \langle u, v \rangle_V$  for any  $u, v \in V$ . Since  $\langle \cdot, \cdot \rangle_V$  is an inner product this implies  $(\psi^* \psi)(u) = u$  for any  $u \in V$ . Thus  $\psi^* \psi = I_V$ .  $\square$

**3.F.33.** Consider the vector  $v = \left(0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right)^T \in \mathbb{R}^3$ . Find an orthogonal operator  $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  such that  $F(v) = e_1$ , where  $e_1 = (1, 0, 0)^T$  is the first vector of the standard basis on  $\mathbb{R}^3$ . Next confirm your result using Sage.

**Solution.** Let  $A$  be the  $3 \times 3$  matrix corresponding to  $F$ , such that  $F(u) = Au$  for all  $u \in \mathbb{R}^3$ . Since  $F$  should be orthogonal,  $A$  satisfies  $AA^T = E$ , where  $E$  is the  $3 \times 3$  identity matrix. Recall that the columns of an orthogonal matrix  $n \times n$  matrix form an orthonormal basis of  $\mathbb{R}^n$ . For the given vector  $v$  we know that  $F(v) = Av = e_1$ , from where we get  $v = A^T e_1$ . Thus  $v$  sits in the first column of  $A^T$ , and we may assume that  $A^T = (v \mid v_1 \mid v_2)$  for some vectors  $v_1, v_2$  on  $\mathbb{R}^3$  that we should specify. These vectors should be orthogonal to  $v$ , i.e.,  $v_1, v_2 \in v^\perp$  and it is easy to see that any vector  $(x_1, x_2, x_3)^T$  on the orthogonal complement  $v^\perp$  should satisfy the equation

$$\frac{\sqrt{2}}{2}x_2 - \frac{\sqrt{2}}{2}x_3 = 0$$

or equivalently,  $x_2 - x_3 = 0$ . Thus, the solution space has the form

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} t \\ s \\ s \end{pmatrix} = t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + s \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} : t, s \in \mathbb{R} \right\}$$

and we deduce that  $v^\perp$  is generated by the vectors  $(1, 0, 0)^T$  and  $(0, 1, 1)^T$ . Obviously, these vectors are orthogonal each other. Setting  $v_1 = (1, 0, 0)^T$  and  $v_2 = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T$  such that  $\|v_1\| = 1 = \|v_2\|$  and  $v_1 \perp v_2$ , we finally obtain

$$A^T = \begin{pmatrix} 0 & 1 & 0 \\ \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{pmatrix} \implies A = \begin{pmatrix} 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}.$$

Using this expression we can verify that  $Av = e_1$ . Consequently, if  $u = (x, y, z)^T$  is an arbitrary vector of  $\mathbb{R}^3$  we have

$$F(u) = Au = \begin{pmatrix} 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2}(y - z) \\ x \\ \frac{\sqrt{2}}{2}(y + z) \end{pmatrix}.$$

As for a confirmation in Sage, use the cell given here:

```
A=matrix([[0, sqrt(2)/2, -sqrt(2)/2], [1, 0, 0], [0, sqrt(2)/2, sqrt(2)/2]])
v=vector([0, sqrt(2)/2, -sqrt(2)/2]); e1=vector([1, 0, 0])
print(A.is_unitary()); print(A*v==e1)
```

□

Let us now analyze an example related to the notion of matrix groups  $G \subset \text{GL}_n(\mathbb{F})$ , where  $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$  and  $\text{GL}_n(\mathbb{F})$  is the general linear group. In this part, to prove the crucial property of “closedness” that characterizes such a group  $G$ , we require a basic understanding of continuous functions, as well as some familiarity with closed subsets of Euclidean space. These concepts are explored in Chapter 5, and our tasks below establish a first link between analysis and group theory. Notably, we have already encountered the concept of “closedness” in the proof of the lemma presented in 3.3.3, and the reader may find it helpful to refer to a similar explanation below.<sup>21</sup>

**3.F.34.** Prove that the orthogonal group  $O(n)$  is a matrix group, that is, a closed subgroup of  $\text{GL}_n(\mathbb{R})$ .

**Solution.** Hints: The fact that  $O(n)$  is a subgroup was discussed in 3.D.18. Consider the map

$$\rho : \text{Mat}_n(\mathbb{R}) \rightarrow \text{Sym}(n, \mathbb{R}), \quad A \mapsto AA^T.$$

Here,  $\text{Sym}(n, \mathbb{R})$  is the set of  $n \times n$  symmetric matrices with real entries. First, establish a linear isomorphism between  $\text{Sym}(n, \mathbb{R})$  and the Euclidean space  $\mathbb{R}^{\frac{1}{2}n(n+1)}$ . Then, deduce that the function  $\rho$  is continuous and moreover that  $O(n) = \rho^{-1}(\{E\})$ , where  $E$  is the  $n \times n$  identity matrix. Since the preimage of a closed set under a continuous map is closed, this shows that  $O(n)$  is closed in  $\text{GL}_n(\mathbb{R})$ , thereby confirming that  $O(n)$  is indeed a matrix group. □

**3.F.35. Matrix of general rotation in  $\mathbb{R}^3$ .** Derive the matrix of a general rotation in  $\mathbb{R}^3$ ,

**Solution.** Consider an arbitrary unit vector  $(x, y, z)^T \in \mathbb{R}^3$ . Rotation in the positive sense by an angle  $\varphi$  about this vector can be written down as a composition of the following rotations (whose matrices we already know):

- rotation  $\mathcal{R}_1$  in the negative sense about the  $z$  axis through the angle with cosine equal to  $x/\sqrt{x^2 + y^2} = x/\sqrt{1 - z^2}$ , that is, with sine  $y/\sqrt{1 - z^2}$ , under which the line with the directional vector  $(x, y, z)^T$  goes over on the line with the directional vector  $(0, y, z)^T$ . The matrix of this rotation has the form

$$R_1 = \begin{pmatrix} x/\sqrt{1 - z^2} & y/\sqrt{1 - z^2} & 0 \\ -y/\sqrt{1 - z^2} & x/\sqrt{1 - z^2} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

- Rotation  $\mathcal{R}_2$  in the positive sense about the  $y$  axis through the angle with cosine  $\sqrt{1 - z^2}$ , that is, with sine  $z$ , under which the line with the directional vector  $(0, y, z)^T$  goes over on the line with the directional vector  $(1, 0, 0)^T$ . The matrix of this rotation is given by

$$R_2 = \begin{pmatrix} \sqrt{1 - z^2} & 0 & z \\ 0 & 1 & 0 \\ -z & 0 & \sqrt{1 - z^2} \end{pmatrix}.$$

- Rotation  $\mathcal{R}_3$  in the positive sense about the  $x$  axis through the angle  $\varphi$ , with the following matrix

$$R_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ 0 & \sin(\varphi) & \cos(\varphi) \end{pmatrix}.$$

- Rotation  $\mathcal{R}_2^{-1}$  with the matrix  $R_2^{-1}$ ,
- Rotation  $\mathcal{R}_1^{-1}$  with the matrix  $R_1^{-1}$ .

<sup>21</sup>Recall that we have an inclusion  $\text{GL}_n(\mathbb{R}) \subset \mathbb{R}^{n^2} \cong \text{Mat}_n(\mathbb{R})$ .

The matrix of the composition of these mappings, that is, the matrix we are looking for, is given by the product of the rotations in the reverse order:

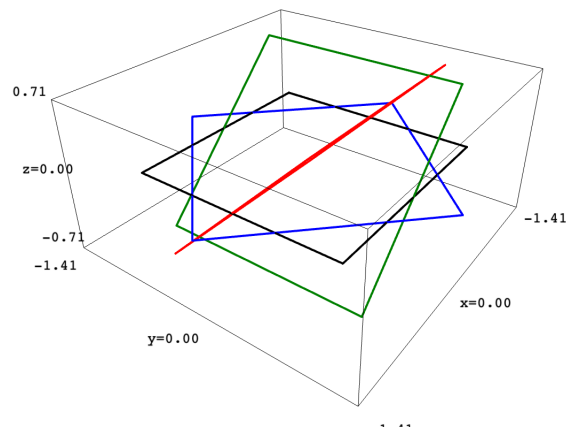
$$R_1^{-1} \cdot R_2^{-1} \cdot R_3 \cdot R_2 \cdot R_1 = \begin{pmatrix} 1 - t + tx^2 & txy - zs & txz + ys \\ yxt + zs & 1 - t + ty^2 & tyz - xs \\ zxt - ys & tzy + xs & 1 - t + tz^2 \end{pmatrix}$$

where  $t = 1 - \cos \varphi$  and  $s = \sin \varphi$ . □

**3.F.36.** Find the matrix of the rotation in the positive sense by the angle  $\pi/3$  about the line passing through the origin with the oriented directional vector  $(1, 1, 0)^T$  under the standard basis  $\mathbb{R}^3$ . ○

**3.F.37.** Using basic plotting functions in Sage, visualize the 2-dimensional square on  $\mathbb{R}^3$ , having vertices the points  $[1, 1, 0]$ ,  $[1, -1, 0]$ ,  $[-1, -1, 0]$  and  $[-1, 1, 0]$ , as it rotates around the  $x$ ,  $y$  and  $z$  axes. Hint: Use the matrices  $R_x$ ,  $R_y$  and  $R_z$  presented in the solution of 3.D.19.

**Solution.** We should define the given square, next, apply the rotation matrices described in ?? and finally, plot the rotated squares. It will be easier to fix the rotation angle, say  $\theta = \pi/4$ , in order to visualize our goal by a figure, as follows:



In this figure, the initial square is shown in black. The square rotated about the  $x$ -axis is shown in red, the square rotated around the  $y$ -axis is in green, and the square rotated around the  $z$ -axis is in blue. Let us present the code and include the necessary explanations within the code block.

```
# Define the vertices of the square in the xy-plane
square_vertices = [(1, 1, 0), (1, -1, 0), (-1, -1, 0), (-1, 1, 0)]
# Define the edges of the square
square_edges = [(square_vertices[0], square_vertices[1]),
                (square_vertices[1], square_vertices[2]),
                (square_vertices[2], square_vertices[3]),
                (square_vertices[3], square_vertices[0])]
# Define the rotation angle in radians theta=pi/4 #45degrees
theta = pi / 4 # 45 degrees
# Rotation matrices
Rx = Matrix([[1, 0, 0], [0, cos(theta), -sin(theta)], [0, sin(theta), cos(theta)]])
Ry = Matrix([[cos(theta), 0, sin(theta)], [0, 1, 0], [-sin(theta), 0, cos(theta)]])
Rz = Matrix([[cos(theta), -sin(theta), 0], [sin(theta), cos(theta), 0], [0, 0, 1]])
# Function to rotate vertices and create plot for each axis
def plot_rotation(rotation_matrix, color, label):
    rotated_edges = []
    for start, end in square_edges:
        start_rotated = rotation_matrix * vector(start)
        end_rotated = rotation_matrix * vector(end)
        rotated_edges.append((start_rotated, end_rotated))
    return sum([line([edge[0], edge[1]], color=color,
                    thickness=2, legend_label=label) for edge in rotated_edges])
```



```

# Plot the original and rotated squares
original_plot = plot_rotation(Matrix.identity(3), "black", "Original")
rotated_x_plot = plot_rotation(Rx, "red", "Rotation around x-axis")
rotated_y_plot = plot_rotation(Ry, "green", "Rotation around y-axis")
rotated_z_plot = plot_rotation(Rz, "blue", "Rotation around z-axis")
# Combine the plots and show them
final_plot = original_plot + rotated_x_plot + rotated_y_plot + rotated_z_plot
final_plot.show(aspect_ratio=1, frame=True, legend=True)
    
```

The most advanced part of our program is the definition of the `plot_rotation` function. This is designed to take a rotation matrix, a colour, and a label as inputs. It then applies the rotation matrix to each edge of the given square, creates lines for these rotated edges, and returns a plot object that can be displayed, making it easier to visualize the effect of the rotation in  $\mathbb{R}^3$ . The command `rotated_edges = []` initializes a list for rotated edges. In particular, this is an empty list that will store the pairs of points (edges) after they have been transformed by the rotation matrix. Next, the syntax

```
for start, end in square_edges:
```

forms a loop over each edge of the square. In particular, the option `square_edges` is a list of tuples, where each tuple represents an edge of the square as a pair of vertex coordinates. The loop iterates over each edge, extracting the start and end points (`start, end`) of the edge. On the other hand, the syntax

```
start_rotated = rotation_matrix * vector(start)
end_rotated = rotation_matrix * vector(end)
```

plays the following role: The first operation multiplies the rotation matrix by a vector created from the start point of the edge. The `vector()` function then converts the coordinate tuple into a vector object, allowing for matrix multiplication. Thus, the syntax `start_rotated` encodes the new vector representing the rotated position of the start point, while, the syntax `end_rotated` encodes the new vector representing the rotated position of the end point. Next, the syntax

```
rotated_edges.append((start_rotated, end_rotated))
```

adds the newly rotated edge, represented as a tuple of the rotated start and end points, to the `rotated_edges` list. Finally, to create line plots for each rotated edge we typed the last line within the `def` environment. This creates a list of line objects, each representing a line plot for an edge. To summarize, the `plot_rotation` function efficiently rotates the square's vertices using a specified rotation matrix, then plots each rotated edge with a given color and label. You can experiment further by changing the rotation angle  $\theta$  from the value we used, i.e.,  $\theta = \pi/4$ .  $\square$

**3.F.38.** Let  $(V, \langle \cdot, \cdot \rangle)$  be a unitary vector space and suppose that  $\varphi : V \rightarrow V$  is a linear mapping having the property  $\varphi^2 = \varphi$ . Prove that there exists a linear subspace of  $W$  such that  $\varphi$  is the orthogonal projection on  $W$ , i.e.,  $\varphi = \text{proj}_W$ , if and only if  $\varphi$  is self-adjoint.

**Solution.** Assume that  $\varphi$  is self adjoint. Then, by linearity and the condition that satisfies  $\varphi$  we see that  $\varphi(u - \varphi(u)) = \varphi(u) - \varphi^2(u) = 0$  for any  $u \in V$ . Thus  $u - \varphi(u) \in \ker(\varphi)$ . Recall now by 3.F.41 that any endomorphism  $\varphi$  on  $V$  satisfies  $\ker(\varphi^*) = (\text{Im}(\varphi))^\perp$  and  $\text{Im}(\varphi^*) = (\ker(\varphi))^\perp$ . By the second relation we get  $\ker(\varphi) = (\text{Im}(\varphi^*))^\perp$  and since  $\varphi$  is self-adjoint this gives  $\ker(\varphi) = (\text{Im}(\varphi))^\perp$ . Thus writing

$$u = \varphi(u) + (u - \varphi(u)),$$

we have  $\varphi(u) \in \text{Im}(\varphi)$  and  $(u - \varphi(u)) \in (\text{Im}(\varphi))^\perp$ . Since  $V = \text{Im}(\varphi) \oplus (\text{Im}(\varphi))^\perp$ , we deduce that  $\varphi(u) = \text{proj}_W u$ , for any  $u \in V$ , where  $W = \text{Im}(\varphi) \subset V$ . This proves the one direction. For the converse, assume that there exists a subspace  $W \subset V$ , such that  $\varphi = \text{proj}_W$ . Let  $W^\perp$  the orthogonal complement of  $W$  with respect to  $\langle \cdot, \cdot \rangle$ . Given arbitrary  $u, v \in V$ , write  $u = u_1 + u_2, v = v_1 + v_2$  with  $u_1, v_1 \in W$  and  $u_2, v_2 \in W^\perp$ , respectively. Then, since  $\varphi = \text{proj}_W$  we have  $\varphi(u) = u_1, \varphi(v) = v_1$ , and thus

$$\langle \varphi(u), v \rangle = \langle u_1, v_1 + v_2 \rangle = \langle u_1, v_1 \rangle = \langle u_1 + u_2, v_1 \rangle = \langle u, v_1 \rangle = \langle u, \varphi(v) \rangle.$$

Thus  $\varphi = \varphi^*$  and  $\varphi$  is self-adjoint. □

**3.F.39.** Show that for any symmetric matrix  $A \in \text{Mat}_n(\mathbb{R})$  the operator  $L_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  defined by  $L_A x = Ax$  ( $x \in \mathbb{R}^n$ ) is self-adjoint. ○

**3.F.40.** For  $A \in \text{Mat}_m(\mathbb{C})$  Hermitian, prove that the linear operator  $L_A : \text{Mat}_{m,n}(\mathbb{C}) \rightarrow \text{Mat}_{m,n}(\mathbb{C})$  defined by  $L_A(B) = AB$  is self-adjoint. ○

**3.F.41.** Let  $f : V \rightarrow V$  be an endomorphism on a (finite-dimensional) inner product space  $(V, \langle \cdot, \cdot \rangle)$ . Prove that  $\ker(f^*) = (\text{Im}(f))^\perp$  and  $\text{Im}(f^*) = (\ker(f))^\perp$ . ○

**3.F.42.** Let  $\varphi_u : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be the linear mapping that reflects  $\mathbb{R}^n$  through the line in the direction of the unit vector  $u \in \mathbb{R}^n$ . Show that its matrix  $[\varphi_u] = 2uu^T - E$  is unitary. ○

**3.F.43.** Let  $A \in \text{Mat}_n(\mathbb{C})$  be a complex matrix. Prove that the product  $A^*A$  has only real eigenvalues. Next suppose that there exists a unitary matrix  $U$  such that  $A = U^*DU$  for some diagonal matrix  $D$ . Show that  $A$  is normal. ○

**3.F.44.** Find the values of the complex parameters  $a, b, c$  such that the matrix  $A$  is a Hermitian matrix, where

$$A = \begin{pmatrix} 1 & a & -i \\ 2 - 2i & 0 & b \\ c & 1 + i & 0 \end{pmatrix}.$$
○

**3.F.45.** Given a Hermitian matrix  $A$  show that its determinant  $\det(A)$  is a real number.

**Solution.** The matrix  $A$  is supposed to be Hermitian, and hence unitary diagonalizable. Thus  $U^*AU = D$  for some unitary matrix  $U$  and diagonal (real) matrix  $D$  (since  $A$  has only real eigenvalues). Then, since  $U^{-1} = U^*$  a direct computation shows that  $A = U^*DU$ . Therefore,

$$\det(A) = \det(U^*) \det(D) \det(U) = \det(U^{-1}) \det(D) \det(U) = \det(D),$$

and our claim follows. □

*So far, we have explored various methods in Sage for applying the Gram-Schmidt orthogonalization process. However, we have not yet discussed the `gram_schmidt()` function, which is only used to orthogonalize the columns of a matrix. When the `gram_schmidt()` method is applied to a matrix, it returns two matrices: a matrix whose columns are mutually orthogonal (but not necessarily of unit length) and an upper triangular matrix that contains the coefficients used in the Gram-Schmidt process to express the original columns as linear combinations of the orthogonalized columns. In 3.F.46 below, we will apply this method and use only the first of the two resulting matrices.*

**3.F.46.** Present an orthogonal diagonalization of the following symmetric matrix  $A = \begin{pmatrix} -\frac{1}{3} & 1 & 0 \\ 1 & -\frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{3} \end{pmatrix}$ . Next implement

the task in Sage using the `gram_schmidt()` function mentioned above.

**Solution.** The characteristic polynomial of  $A$  is given by

$$\chi_A(\lambda) = \begin{vmatrix} -\frac{1}{3} - \lambda & 1 & 0 \\ 1 & -\frac{1}{3} - \lambda & 0 \\ 0 & 0 & -\frac{1}{3} - \lambda \end{vmatrix} = -(\lambda + \frac{1}{3})(\lambda^2 + \frac{2}{3}\lambda - \frac{8}{9}) = -(\lambda + \frac{1}{3})(\lambda + \frac{4}{3})(\lambda - \frac{2}{3}).$$

Thus, the eigenvalues of  $A$  are given by  $\lambda_1 = \frac{2}{3}$ ,  $\lambda_2 = -\frac{2}{3}$  and  $\lambda_3 = -\frac{4}{3}$ , all with algebraic multiplicity one. The geometric multiplicity of each  $\lambda_i$  is also one, for all  $i = 1, 2, 3$ , and hence the matrix  $A$  is diagonalizable. Let us find the corresponding eigenvectors.

- For  $\lambda_1 = \frac{2}{3}$  we need to solve the matrix equation  $(A - \frac{2}{3}E)u = 0$  for some vector  $u = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ . We see that  $A - \frac{2}{3}E = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$  and the corresponding linear system has solution space  $\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} t \\ t \\ 0 \end{pmatrix} : t \in \mathbb{R} \right\}$ . Hence,

the eigenvectors corresponding to  $\lambda_1$  are multiples of the vector  $u_1 = (1, 1, 0)^T$ .

- For  $\lambda_2 = -\frac{1}{3}$  we have the matrix equation  $(A + \frac{1}{3}E)u = 0$  for some vector  $u = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ , where  $A + \frac{1}{3}E = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ . The solution space of the corresponding linear system has the form  $\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ s \end{pmatrix} : s \in \mathbb{R} \right\}$ . Hence, the

eigenvectors corresponding to  $\lambda_2$  are multiples of the vector  $u_2 = (0, 0, 1)^T$ .

- For  $\lambda_3 = -\frac{4}{3}$  we have the matrix equation  $(A + \frac{4}{3}E)u = 0$  for some vector  $u = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ , where  $A + \frac{4}{3}E =$

$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ . The solution space of the corresponding linear system has the form  $\left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} r \\ -r \\ 0 \end{pmatrix} : r \in \mathbb{R} \right\}$ . Hence, the eigenvectors corresponding to  $\lambda_3$  are multiples of the vector  $u_3 = (1, -1, 0)^T$ .

In Sage we can confirm these results as follows:

```

A=matrix(QQ, [[-1/3, 1, 0], [1, -1/3, 0], [0, 0, -1/3]])
chi_A=A.characteristic_polynomial()
print(chi_A.factor())
print(A.eigenvalues())
show(A.eigenmatrix_right())
D, Pein=A.eigenmatrix_right()
show(D, Pein)
    
```

Executing this cell Sage prints out the characteristic polynomial of  $A$ , its eigenvalues, the diagonal matrix  $D = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$  and the matrix whose columns are the eigenvectors  $u_1, u_2, u_3$  (the latter is denoted by `Pein` inside the program).

Next, we see that  $u_1 \perp u_2$ ,  $u_1 \perp u_3$  and  $u_2 \perp u_3$ :

$$\langle u_1, u_2 \rangle = u_2^T u_1 = 0, \quad \langle u_1, u_3 \rangle = u_3^T u_1 = 0, \quad \langle u_2, u_3 \rangle = u_3^T u_2 = 0.$$

Moreover, we compute  $\|u_1\| = \sqrt{2} = \|u_3\|$ , while  $\|u_2\| = 1$ . Thus, the eigenvectors given below are orthonormal:

$$\hat{u}_1 = \frac{u_1}{\|u_1\|} = \left( \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0 \right)^T, \quad \hat{u}_2 = u_2 = (0, 0, 1)^T, \quad \hat{u}_3 = \frac{u_3}{\|u_3\|} = \left( \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0 \right)^T.$$

We are now ready to present the matrix  $P$ :

$$P = (\hat{u}_1 \mid \hat{u}_2 \mid \hat{u}_3) = \begin{pmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \\ 0 & 1 & 0 \end{pmatrix}.$$

By adding the following cell to the previous block, we can quickly verify that  $P$  is orthogonal, and furthermore that

$$P^{-1}AP = P^TAP = D = \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ 0 & -\frac{2}{3} & 0 \\ 0 & 0 & -\frac{4}{3} \end{pmatrix}.$$

```

P=matrix([[sqrt(2)/2, 0, sqrt(2)/2], [sqrt(2)/2, 0, -sqrt(2)/2], [0, 1, 0]])
print(P.is_unitary())
print(D==P.inverse()*A*P)
    
```

Let us now implement the solution of the task using the `gram_schmidt` method, as suggested in the statement. As mentioned earlier, we will only utilize the first matrix returned by this method. To normalize the column vectors of a given matrix, we will define another function, which we may call `normalize_col`. This takes a matrix  $M$  as input and returns another matrix, with each column normalized. It works as follows:

- It iterates through each column vector  $u$  of the matrix  $M$ .
- Then, each vector  $u$  is divided by its norm to normalize it.
- Finally the normalized columns are recombined into a new matrix, using the `column_matrix()` command.

For our task we also need the matrices `D` and `Pein`, as introduced in the initial block via the `eigenmatrix_right` command. Our program has the following form:

```
A=matrix(QQ, [[-1/3, 1, 0], [1, -1/3, 0], [0, 0, -1/3]])
D, Pein=A.eigenmatrix_right()
# store the transpose of the matrix whose columns are the unnormalized eigenvectors of A as Q
Q=Pein.T
# orthogonalize the columns of Q
Q.gram_schmidt()[0]
# take the transpose of the matrix obtained by Gram-Schmidt
R=Q.gram_schmidt()[0].T
show(R)
#construct the function that normalizes the column vectors of a matrix M
def normalize_col(M):
    return column_matrix([v/norm(v) for v in M.columns()])
```

```
#normalize the columns of the matrix R and store the resulting matrix as P
P=normalize_col(R)
show(P)
print(P.is_unitary())
print(P.inverse()*A*P==D)
```

Running this program, Sage displays the matrix  $P$  as shown below, along with the message “True” printed twice, verifying that  $P$  is an orthogonal matrix that satisfies the relation  $P^TAP = D$ :

$$\begin{pmatrix} \frac{1}{2}\sqrt{2} & 0 & \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & 0 & -\frac{1}{2}\sqrt{2} \\ 0 & 1 & 0 \end{pmatrix}.$$

□

**3.F.47.** Using the program in Sage constructed in 3.F.46, present the orthogonal diagonalization of the following matrices:

$$A = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 0 & 4 \\ 2 & 4 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -4 & 4 & 4 \\ 4 & -4 & 4 \\ 4 & 4 & -4 \end{pmatrix}.$$

**Solution.** For the matrix  $A$  the program has the form

```
A=matrix(QQ, [[2, 2, 2], [2, 0, 4], [2, 4, 0]])
show(A.eigenmatrix_right())
D, Pein=A.eigenmatrix_right()
show(D,Pein)
Q=Pein.T
R=Q.gram_schmidt()[0].T
def normalize_columns(M):
    return column_matrix([v/norm(v) for v in M.columns()])
P=normalize_columns(R);show(P)
print(P*D*P.T==A); print(P.is_unitary())
```

In this case we get the orthogonal matrix  $P = \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{2}{3}\sqrt{\frac{3}{2}} & 0 \\ \frac{\sqrt{3}}{3} & -\frac{1}{3}\sqrt{\frac{3}{2}} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} & -\frac{1}{3}\sqrt{\frac{3}{2}} & -\frac{\sqrt{2}}{2} \end{pmatrix}$ , such that

$$P^TAP = \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ \frac{2}{3}\sqrt{\frac{3}{2}} & -\frac{1}{3}\sqrt{\frac{3}{2}} & -\frac{1}{3}\sqrt{\frac{3}{2}} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} 2 & 2 & 2 \\ 2 & 0 & 4 \\ 2 & 4 & 0 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{2}{3}\sqrt{\frac{3}{2}} & 0 \\ \frac{\sqrt{3}}{3} & -\frac{1}{3}\sqrt{\frac{3}{2}} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} & -\frac{1}{3}\sqrt{\frac{3}{2}} & -\frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -4 \end{pmatrix} = D,$$

where  $\lambda_1 = 6$ ,  $\lambda_2 = 0$  and  $\lambda_3 = -4$  are the eigenvalues of  $A$ .

Similarly for the matrix  $B$ :

```
B=matrix(QQ, [[-4, 4, 4], [4, -4, 4], [4, 4, -4]])
show(B.eigenmatrix_right())
D, Pein=B.eigenmatrix_right()
show(D,Pein)
Q=Pein.T
show(Q.gram_schmidt()[0])
R=Q.gram_schmidt()[0].T
show(R)
def normalize_columns(M):
    return column_matrix([v/norm(v) for v in M.columns()])
P=normalize_columns(R)
show(P)
print(P*D*P.T==B)
print(P.is_unitary())
```

Running this block Sage prints out the orthogonal matrix  $P = \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} & -\frac{1}{3}\sqrt{\frac{3}{2}} \\ \frac{\sqrt{3}}{3} & 0 & \frac{2}{3}\sqrt{\frac{3}{2}} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{2}}{2} & -\frac{1}{3}\sqrt{\frac{3}{2}} \end{pmatrix}$ , such that

$$P^T B P = \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \\ -\frac{1}{3}\sqrt{\frac{3}{2}} & \frac{2}{3}\sqrt{\frac{3}{2}} & -\frac{1}{3}\sqrt{\frac{3}{2}} \end{pmatrix} \begin{pmatrix} -4 & 4 & 4 \\ 4 & -4 & 4 \\ 4 & 4 & -4 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{2}}{2} & -\frac{1}{3}\sqrt{\frac{3}{2}} \\ \frac{\sqrt{3}}{3} & 0 & \frac{2}{3}\sqrt{\frac{3}{2}} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{2}}{2} & -\frac{1}{3}\sqrt{\frac{3}{2}} \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & -8 & 0 \\ 0 & 0 & -8 \end{pmatrix} = D,$$

where  $\lambda_1 = 4$  and  $\lambda_2 = -8$  are the eigenvalues of  $B$ , the second with an algebraic multiplicity of two.  $\square$

**3.F.48.** Present a unitary diagonalization of the matrix  $B = \begin{pmatrix} 0 & i & 1 \\ -i & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ .

**Solution.** The matrix  $B$  is Hermitian, so it has real eigenvalues given by  $\lambda_1 = 0$  and  $\lambda_{2,3} = \pm\sqrt{2}$ . The corresponding eigenvectors have the form

$$u_1 = (0, i, 1)^T, \quad u_2 = (\sqrt{2}, -i, 1)^T, \quad u_3 = (-\sqrt{2}, -i, 1)^T.$$

By running in a sequence the following commands in Sage, one can verify that  $\langle u_i, u_j \rangle = 0$  for any  $1 \leq i \neq j \leq 3$ , and  $\|u_1\| = \sqrt{2}$ ,  $\|u_2\| = \|u_3\| = 2$ .

```
u1=vector(QQbar, [0, I, 1])
u2=vector(QQbar, [sqrt(2), -I, 1])
u3=vector(QQbar, [-sqrt(2), -I, 1])
u2.hermitian_inner_product(u1)
u1.hermitian_inner_product(u3)
u3.hermitian_inner_product(u2)
u1.norm(); u2.norm(); u3.norm()
```

The normalized eigenvectors  $\hat{u}_i = u_i/\|u_i\|$  for  $i = 1, 2, 3$  form the columns of the matrix  $U$ ,

$$U = \begin{pmatrix} 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{i\sqrt{2}}{2} & -\frac{i}{2} & -\frac{i}{2} \\ \frac{\sqrt{2}}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

We can verify that  $U$  is unitary; for example in Sage you can type

```
U=matrix(QQbar, [[0, sqrt(2)/2, -sqrt(2)/2],
[(sqrt(2)/2)*I, (-1/2)*I, (-1/2)*I],
[sqrt(2)/2, 1/2, 1/2]])
U.is_unitary()
```

which returns: True. This diagonalizes the given matrix  $B$  with  $D = \text{diag}(0, \sqrt{2}, -\sqrt{2})$ . Recall by 3.D.38 that a verification in Sage of this final claim relies on the block

```
U=matrix(QQbar, [[0, sqrt(2)/2, -sqrt(2)/2],
[(sqrt(2)/2)*I, (-1/2)*I, (-1/2)*I],
[sqrt(2)/2, 1/2, 1/2]])
A = matrix(QQbar, [[0, I, 1], [-I, 0, 0], [1, 0, 0]])
D = matrix(QQbar, [[0, 0, 0], [0, sqrt(2), 0], [0, 0, -sqrt(2)]])
U.conjugate_transpose()*A*U == D
```

Run this block yourself to see the output generated by Sage.  $\square$

**3.F.49.** Adapt the method of orthogonal diagonalization presented in 3.F.46, to implement the unitary diagonalization of the matrix  $B$  given in 3.F.48 using Sage.

**Solution.** We only need to make a few adjustments to the program presented in 3.F.46 for it to work with unitary diagonalization. In particular, since unitary diagonalization deals with complex matrices we need to replace carefully the transposes by conjugate transposes. Also, to achieve a more readable display of the printed matrices by Sage, it is useful to pass the matrix to a low-precision ring before printing it. This adjustment will not effect the original matrices and the final results, only their display. Below is the our program:

```
CC20 = ComplexField(prec=20)
MatPrint = MatrixSpace(CC20, 3, 3)
B=matrix(QQbar, [[0, i, 1], [-i, 0, 0], [1, 0, 0]])
show(MatPrint(B))
print(B==B.H)
chi_B=B.characteristic_polynomial()
```

```

print(chi_B.factor())
print(B.eigenvalues())
D, Pein=B.eigenmatrix_right()
show(MatPrint(D))
show(MatPrint(Pein))
Q=Pein.H #show(MatPrint(Q))
q=Q.gram_schmidt()[0]
R=Q.gram_schmidt()[0].H
def normalize_columns(M):
    return column_matrix([v/norm(v) for v in M.columns()])

P=normalize_columns(R)
show(MatPrint(P))
print(P*D*P.H==B); print(P.is_unitary())
    
```

Sage's prints the solution matrix  $P$  as follows

$$P = \begin{pmatrix} 0.70711 & 9.5829 \times 10^{-21}i & 0.70711 \\ -0.50000i & 0.70711 & 0.50000i \\ 0.50000 & -0.70711i & -0.50000 \end{pmatrix} = \begin{pmatrix} 0.70711 & 0 & 0.70711 \\ -0.50000i & 0.70711 & 0.50000i \\ 0.50000 & -0.70711i & -0.50000 \end{pmatrix}.$$

This matrix slightly differs from our solution matrix  $P$  in 3.F.48, since for example Sage enumerates the eigenvalues differently and uses different eigenvectors (recall that  $P$  is not uniquely determined). However, Sage verifies that  $P$  is unitary and satisfies the unitary diagonalization condition. To understand the importance of using the low-precision ring before printing our matrices, try running the program in Sage without including the initial lines related to matrix display. Then, compare the corresponding output with the one presented above.  $\square$

Next we will explore several tasks involving the bracket of square matrices of the same size, also known as the commutator  $[\cdot, \cdot]$ , defined as  $[A, B] := AB - BA$ . The commutator is a fundamental operation in linear algebra and plays a crucial role in various branches of mathematics and physics, particularly in the study of the so called “Lie algebras”. When  $[A, B] = 0$ , the matrices commute in the sense that  $AB = BA$ . Hence, the commutator  $[A, B]$  measures the extend to which two matrices  $A, B$  fail to commute.

The commutator is a central concept in the theory of Lie algebras, which are algebraic structures used to study symmetry, particularly in the context of continuous transformations. A “Lie algebra” is a vector space  $V$  over a field  $\mathbb{F}$ , equipped with a bilinear operation  $[\cdot, \cdot] : V \times V \rightarrow V$ , known as the “Lie bracket”. This is skew-symmetric and satisfies the so called Jacobi identity

$$\mathfrak{S}_{X,Y,Z}[X, [Y, Z]] = 0,$$

where  $\mathfrak{S}_{X,Y,Z}$  is the cyclic sum over the elements  $X, Y, Z \in V$ .

We should mention that while the commutator of matrices is a primary example, the concept of a Lie algebra extends far beyond matrices, encompassing a broad range of algebraic structures.

**3.F.50.** For  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{F} = \mathbb{C}$ , prove that the vector space of square matrices  $\text{Mat}_n(\mathbb{F})$  endowed with the commutator  $[A, B] := AB - BA$  has the structure of a (finite-dimensional) Lie algebra.

**Solution.** It suffices to prove that the commutator is a bilinear mapping  $[\cdot, \cdot] : \text{Mat}_n(\mathbb{F}) \times \text{Mat}_n(\mathbb{F}) \rightarrow \text{Mat}_n(\mathbb{F})$ , which is skew-symmetric and satisfies the Jacobi identity. Consider arbitrary matrices  $A_1, A_2 \in \text{Mat}_n(\mathbb{F})$  and scalars  $\lambda, \mu \in \mathbb{F}$ . Then, for any  $B \in \text{Mat}_n(\mathbb{F})$  we see that

$$\begin{aligned} [\lambda A_1 + \mu A_2, B] &= (\lambda A_1 + \mu A_2)B - B(\lambda A_1 + \mu A_2) = \lambda A_1 B + \mu A_2 B - \lambda B A_1 - \mu B A_2 \\ &= \lambda(A_1 B - B A_1) + \mu(A_2 B - B A_2) = \lambda[A_1, B] + \mu[A_2, B]. \end{aligned}$$

This prove linearity in the first argument and in a similar way we obtain  $[A, \lambda B_1 + \mu B_2] = \lambda[A, B_1] + \mu[A, B_2]$ , for all  $A, B_1, B_2 \in \text{Mat}_n(\mathbb{F})$  and  $\lambda, \mu \in \mathbb{F}$ . Since the bracket  $[\cdot, \cdot]$  is linear also in the second argument, is a bilinear mapping. Skew-symmetry is obvious:  $[B, A] = BA - AB = -(AB - BA) = -[A, B]$ , for any  $A, B \in \text{Mat}_n(\mathbb{F})$ . Finally, for the Jacobi identity, let  $A, B, C$  be some arbitrary  $n \times n$  matrices over  $\mathbb{F}$ . Then, we see that

$$\begin{aligned} [A, [B, C]] &= A(BC - CB) - (BC - CB)A = ABC - ACB - BCA + CBA, \\ [B, [C, A]] &= B(CA - AC) - (CA - AC)B = BCA - BAC - CAB + ACB, \\ [C, [A, B]] &= C(AB - BA) - (AB - BA)C = CAB - CBA - ABC + BAC, \end{aligned}$$

and the Jacobi identity  $[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0$  is now direct.



The Lie algebra  $(\text{Mat}_n(\mathbb{R}), [ , ])$ , also denoted by  $\mathfrak{gl}(n, \mathbb{R})$  or  $\mathfrak{gl}_n(\mathbb{R})$ , underpins a wide range of applications in science, engineering, and technology. Its significance stems from its ability to model and solve real-world problems that involve linear transformations and systems. Note that in a similar way one can prove that  $(\text{Mat}_n(\mathbb{C}), [ , ])$  is a Lie algebra over  $\mathbb{C}$ .  $\square$

**3.F.51.** The dimension of a Lie algebra  $(V, [ , ]_V)$  over a field  $\mathbb{F}$ , is the dimension of the vector space  $V$  over  $\mathbb{F}$ . What is the dimension of the Lie algebra  $(\text{Mat}_n(\mathbb{R}), [ , ])$  described above?  $\circ$

**3.F.52.** Show that the set of matrices  $M = \left\{ \begin{pmatrix} a & b \\ 0 & -a \end{pmatrix} : a, b \in \mathbb{R} \right\}$  forms a Lie algebra under matrix commutation.

**Solution.** The given set of matrices is a linear subspace of the vector space  $\text{Mat}_2(\mathbb{R})$ , the latter consisting of  $2 \times 2$  real matrices. This is because  $M$  contains the zero matrix (it occurs for  $a = b = 0$ ), and is closed under addition and scalar multiplication:

$$A + B = \begin{pmatrix} a_1 & b_1 \\ 0 & -a_1 \end{pmatrix} + \begin{pmatrix} a_2 & b_2 \\ 0 & -a_2 \end{pmatrix} = \begin{pmatrix} a_1 + a_2 & b_1 + b_2 \\ 0 & -(a_1 + a_2) \end{pmatrix} \in M, \quad \lambda A = \begin{pmatrix} \lambda a & \lambda b \\ 0 & -\lambda a \end{pmatrix} \in M,$$

for all  $A, B \in M$  and  $\lambda \in \mathbb{R}$ . Next we should prove that the commutator  $[A, B]$  for  $A, B \in M$  satisfies the properties of a Lie bracket, as defined above and hence  $M$  is a 2-dimensional Lie algebra. An easier way relies on proving that  $M$  is a ‘‘Lie subalgebra’’ of the Lie algebra  $(\text{Mat}_2(\mathbb{R}), [ , ])$ . A *Lie subalgebra* of a Lie algebra  $(V, [ , ]_V)$  is a subspace  $W \subset V$  which is closed under the corresponding Lie bracket operation, i.e.,  $[W, W]_V \subset W$ . It is easy to see that a Lie subalgebra is a Lie algebra itself. For two elements  $A, B \in M$  we see that

$$\begin{aligned} [A, B] &= AB - BA = \begin{pmatrix} a_1 & b_1 \\ 0 & -a_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ 0 & -a_2 \end{pmatrix} - \begin{pmatrix} a_2 & b_2 \\ 0 & -a_2 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ 0 & -a_1 \end{pmatrix} \\ &= \begin{pmatrix} a_1 a_2 & a_1 b_2 - b_1 a_2 \\ 0 & a_1 a_2 \end{pmatrix} - \begin{pmatrix} a_2 a_1 & a_2 b_1 - b_2 a_1 \\ 0 & a_2 a_1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in M. \end{aligned}$$

Thus  $M$  is closed under the commutator, making it a Lie subalgebra of  $\text{Mat}_2(\mathbb{R})$  and hence a Lie algebra in its own right. Since  $[A, B] = 0$  for any  $A, B \in M$ , this Lie algebra is an example of an *Abelian Lie algebra*, a special type of Lie algebras where all Lie brackets are zero.  $\square$

**3.F.53.** Let us denote by  $\mathfrak{so}(3)$  the set of skew-symmetric  $3 \times 3$  matrices.

- Prove that  $\mathfrak{so}(3)$  is a vector subspace of  $\text{Mat}_3(\mathbb{R})$  and compute its dimension.
- Prove that  $\mathfrak{so}(3)$  endowed with the matrix commutator  $[ , ]$  is a Lie subalgebra of  $(\text{Mat}_3(\mathbb{R}), [ , ])$ . Next compute the brackets of the basis elements and use Sage to verify the result.
- Show that  $\mathbb{R}^3$  endowed with the cross product  $\times$  as a Lie bracket, forms another 3-dimensional Lie algebra.
- Establish an linear isomorphism  $\varphi$  between  $\mathbb{R}^3$  and  $\mathfrak{so}(3)$  that preserves the Lie algebra structures, i.e.,  $\varphi(u \times v) = [\varphi(u), \varphi(v)]$ .<sup>22</sup>

**Solution.** (i) An element in  $\mathfrak{so}(3)$  has the form  $\begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$  for some  $x_1, x_2, x_3 \in \mathbb{R}$ . Thus it is easy to see that

$A + B \in \mathfrak{so}(3)$  and  $\lambda A \in \mathfrak{so}(3)$  for any two matrices  $A, B \in \mathfrak{so}(3)$  and scalar  $\lambda \in \mathbb{R}$ . Hence  $\mathfrak{so}(3)$  is a vector subspace of  $\text{Mat}_3(\mathbb{R})$  with  $\dim \mathfrak{so}(3) = 3$ , as a basis is given by the matrices

$$E_{12} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad E_{23} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad E_{31} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}.$$

(ii) Let us prove that  $\mathfrak{so}(3)$  is closed under the matrix commutator. Let  $A, B \in \mathfrak{so}(3)$  be skew-symmetric matrices, i.e.,  $A^T = -A$  and  $B^T = -B$ . Then we see that

$$([A, B])^T = (AB - BA)^T = B^T A^T - A^T B^T = (-B)(-A) - (-A)(-B) = BA - AB = -[A, B],$$

which implies that  $[A, B] \in \mathfrak{so}(3)$ . Thus  $[\mathfrak{so}(3), \mathfrak{so}(3)] \subset \mathfrak{so}(3)$  and  $\mathfrak{so}(3)$  is Lie subalgebra of  $\text{Mat}_3(\mathbb{R})$ , and hence a Lie algebra itself. Alternatively, one may compute the Lie brackets of the basis elements, and prove that they belong to  $\mathfrak{so}(3)$ ,

$$[E_{12}, E_{23}] = E_{31} \in \mathfrak{so}(3), \quad [E_{12}, E_{31}] = -E_{23} \in \mathfrak{so}(3), \quad [E_{23}, E_{31}] = E_{12} \in \mathfrak{so}(3).$$

To verify the brackets in Sage you may use the following block (see also 2.E.66).

<sup>22</sup>Such isomorphisms are called *Lie algebra isomorphisms*.

```
E12=matrix([[0, -1, 0], [1, 0, 0], [0, 0, 0]])  
E23=matrix([[0, 0, 0], [0, 0, -1], [0, 1, 0]])  
E31=matrix([[0, 0, 1], [0, 0, 0], [-1, 0, 0]])
```

```

show("\n\nThe matrix E12 has the form:", E12)
show("\n\nThe matrix E23 has the form:", E23)
show("\n\nThe matrix E31 has the form:", E31)
def Lie_bracket(A, B):
    return A*B-B*A
show("\n\nThe commutator [E12, E23] equals to:", Lie_bracket(E12, E23))
show("\n\nThe commutator [E12, E31] equals to:", Lie_bracket(E12, E31))
show("\n\nThe commutator [E23, E31] equals to:", Lie_bracket(E23, E31))
print(Lie_bracket(E12, E23)==E31)
print(Lie_bracket(E12, E31)==-E23)
print(Lie_bracket(E23, E31)==E12)
    
```

(iii) Recall that the cross product  $v \times w$  of two vectors  $v = (v_1, v_2, v_3)^T$  and  $w = (w_1, w_2, w_3)^T$  on  $\mathbb{R}^3$  is the vector

$$v \times w = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = \begin{pmatrix} v_2 w_3 - w_2 v_3 \\ v_3 w_1 - w_3 v_1 \\ v_1 w_2 - w_1 v_2 \end{pmatrix} \in \mathbb{R}^3.$$

Hence  $\mathbb{R}^3$  is closed with respect to the operation  $[v, w]_{\mathbb{R}^3} := v \times w$ . This is by definition bilinear and skew-symmetric,  $v \times w = -w \times v$  for all  $v, w \in \mathbb{R}^3$ . For the Jacobi identity recall that  $u \times (v \times w) = (u \cdot w)v - (u \cdot v)w$ . Thus

$$\begin{aligned} [u, [v, w]_{\mathbb{R}^3}]_{\mathbb{R}^3} + [v, [w, u]_{\mathbb{R}^3}]_{\mathbb{R}^3} + [w, [u, v]_{\mathbb{R}^3}]_{\mathbb{R}^3} &= u \times (v \times w) + v \times (w \times u) + w \times (u \times v) \\ &= (u \cdot w)v - (u \cdot v)w + (v \cdot u)w - (v \cdot w)u + (w \cdot v)u - (w \cdot u)v \\ &= ((u \cdot w) - (w \cdot u))v - ((u \cdot v) - (v \cdot u))w + ((w \cdot v) - (v \cdot w))u \\ &= 0. \end{aligned}$$

Thus  $(\mathbb{R}^3, [\cdot, \cdot]_{\mathbb{R}^3})$  is also Lie algebra.

(iv) The map  $\varphi: \mathbb{R}^3 \rightarrow \mathfrak{so}(3)$  with

$$u = (x_1, x_2, x_3)^T \mapsto A_u := \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$$

is clearly a vector space isomorphism. To show that is a Lie algebra isomorphism it remains to prove that  $\varphi([u, v]_{\mathbb{R}^3}) = [\varphi(u), \varphi(v)]$  or equivalently,  $A_{u \times v} = [A_u, A_v]$ , which we leave as an exercise.  $\square$

**3.F.54.** Consider Hermitian matrices  $A, B, C$  satisfying  $[A, C] = [B, C] = 0$  and  $[A, B] \neq 0$ . Prove that at least one eigenspace of the matrix  $C$  has dimension  $> 1$ .

**Solution.** We prove it by contradiction. Assume that all eigenspaces of the matrix  $C$  are 1-dimensional. Then, any vector  $u$  can be written as  $u = \sum_k \langle u, u_k \rangle u_k = \sum_k c_k u_k$ , where  $u_k$  are the linearly independent eigenvectors of  $C$ , associated with the eigenvalue  $\lambda_k$ . Now a computation shows that

$$0 = [A, C]u_k = ACu_k - CAu_k = \lambda_k Au_k - C(Au_k).$$

Therefore  $Au_k$  is also an eigenvector of the matrix  $C$  corresponding to eigenvalue  $\lambda_k$ . Then we get  $Au_k = \lambda_k^A u_k$  for some number  $\lambda_k^A$ . Similarly,  $Bu_k = \lambda_k^B u_k$  for some number  $\lambda_k^B$ . Then for the commutator of matrices  $A$  and  $B$  one computes:

$$[A, B]u_k = ABu_k - BAu_k = \lambda_k^A \lambda_k^B u_k - \lambda_k^B \lambda_k^A u_k = 0$$

and hence

$$[A, B]u = [A, B] \sum_k c_k u_k - \sum_k c_k [A, B]u_k = 0.$$

This final relation implies  $[A, B] = 0$ , a contradiction.  $\square$

**3.F.55.** Given the matrix  $A = \begin{pmatrix} 0 & -2 & 2 \\ -2 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}$ , compute the trace of the matrix

$$\exp(A) := \sum_{k=0}^{\infty} \frac{1}{k!} A^k = I_2 + A + \frac{A^2}{2} + \dots$$

Hint: Use the eigenvalues of  $A$ .

**Solution.** Recall that trace of a matrix  $A \in \text{Mat}_n(\mathbb{F})$ , for  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{F} = \mathbb{C}$  equals to the sum of the eigenvalues of  $A$ . The given matrix is symmetric and hence its eigenvalues are real. An application of the following block in Sage gives  $\lambda_1 = -4$ , with multiplicity 1 and  $\lambda_2 = 2$  with multiplicity 2:

```
A=matrix(QQ, 3, 3, [[0, -2, 2], [-2, 0, 2], [2, 2, 0]])
A.eigenvalues ()
```

Thus, the eigenvalues of  $\exp(A)$  are  $e^{-4}$  and  $e^2$ , the second with multiplicity 2 again. It follows that  $\text{tr}(\exp(A)) =$  sum of eigenvalues of  $\exp(A) = e^{-4} + 2e$ .  $\square$

**3.F.56.** Show that if  $H$  is a Hermitian matrix, then  $U = \exp(iH) = \sum_{n=0}^{\infty} \frac{1}{n!} (iH)^n$  is a unitary matrix. Next compute its determinant.

**Solution.** From the definition of  $\exp$  we can show that  $\exp(A + B) = \exp(A)\exp(B)$  just as with the exponential mapping in the domain of real numbers. Because  $(u + v)^* = u^* + v^*$  and  $(cv)^* = \bar{c}v^*$ , we obtain

$$U^* = \left( \sum_{n=0}^{\infty} \frac{1}{n!} (iH)^n \right)^* = \sum_{n=0}^{\infty} \frac{1}{n!} (-iH^*)^n.$$

Since  $H^* = H$ , we finally see that

$$U^* = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!} (iH)^n = \exp(-iH).$$

Thus  $U^*U = \exp(iH)\exp(-iH) = \exp(0) = 1$ , and  $\det(U) = \exp(\text{tr}(iH))$ .  $\square$

*We will now continue with additional tasks related to generalized eigenspaces and the Jordan canonical form.*

**3.F.57.** Determine the algebraic and geometric multiplicities of the eigenvalues for the matrices below, and use Sage to validate your results:

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}, \quad C = \begin{pmatrix} 4 & 0 & 1 \\ 2 & 3 & 2 \\ 1 & 0 & 4 \end{pmatrix}.$$

**Solution.** The matrices  $A$  and  $B$  are upper triangular having the same diagonal entries. Thus  $A, B$  have the same eigenvalues,  $|A - \lambda E| = |B - \lambda E| = \lambda^3 - 5\lambda^2 + 7\lambda - 3 = (\lambda - 1)^2(\lambda - 3)$  and we get  $\lambda_1 = 1$  with multiplicity 2,  $\alpha_A(\lambda_1) = 2 = \alpha_B(\lambda_1)$  and  $\lambda_2 = 3$  with multiplicity 1,  $\alpha_A(\lambda_2) = 1 = \alpha_B(\lambda_2)$ . Using Sage we can obtain the same conclusion as follows:

```
A=matrix(QQbar, [[1, 1, 2], [0, 1, 2], [0, 0, 3]])
A.characteristic_polynomial ()
B=matrix(QQbar, [[1, 0, 2], [0, 1, 2], [0, 0, 3]])
B.characteristic_polynomial ()
A.characteristic_polynomial () == B.characteristic_polynomial ()
```

Recall also that in Sage the command `A.fcp('t')` factors the characteristic polynomial corresponding to a matrix. For the matrix  $C$ , the following cell

```
C=matrix(QQbar, [[4, 0, 1], [2, 3, 2], [1, 0, 4]])
C.fcp('x')
C.eigenvalues ()
```

gives the characteristic polynomial of  $C$  as well as its eigenvalues:  $\lambda_1 = 5$  with  $\alpha_C(\lambda_1) = 1$  and  $\lambda_2 = 1$  with  $\alpha_C(\lambda_2) = 2$ . Let us now compute the geometric multiplicities,

$$\gamma_A(\lambda) = \dim_{\mathbb{C}}(\ker(A - \lambda E)) = n - \text{rank}(A - \lambda E).$$

Here  $A$  is a  $n \times n$  matrix and  $\lambda$  is an eigenvalue of  $A$ .

Thus, for the given matrix  $A$  we get  $A - E = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{pmatrix}$ . In particular,  $A - E$  consists of 2 independent column

vectors. Thus  $\text{rank}(A - E) = 2$  and  $\gamma_A(1) = 3 - 2 = 1$ . Here is a confirmation using Sage:

```
A=matrix(QQbar, [[1, 1, 2], [0, 1, 2], [0, 0, 3]])
E =matrix(QQbar, [[1, 0, 0], [0, 1, 0], [0, 0, 1]])
(A-E).rank ()
```

Or we can type

```
A=matrix(QQbar, [[1, 1, 2], [0, 1, 2], [0, 0, 3]])
A.eigenspaces_right ()
```

The second block returns the eigenvalues of  $A$  and their eigenspaces:

```
(3, Vector space of degree 3 and dimension 1 over Algebraic Field
User basis matrix:
[ 1.0000000000000000? 0.6666666666666667? 0.6666666666666667?]),
(1, Vector space of degree 3 and dimension 1 over Algebraic Field
User basis matrix:
[1 0 0])
```

In a similar way you can verify that  $\gamma_B(1) = 2$ ,  $\gamma_A(3) = 1 = \gamma_B(3)$ ,  $\gamma_C(5) = 1$  and  $\gamma_C(3) = 2$ . □

**3.F.58.** Find the Jordan canonical form of the matrices  $A = \begin{pmatrix} -1 & 1 \\ -6 & 4 \end{pmatrix}$  and  $B = \begin{pmatrix} -1 & 1 \\ -4 & 3 \end{pmatrix}$ . Additionally, provide the geometric interpretation of the Jordan canonical form decomposition corresponding to these matrices.

**Solution.** The eigenvalues of  $A$  are real, given by  $\lambda_1 = 1$  and  $\lambda_2 = 2$ . Since the size of  $A$  is  $2 \times 2$  and it has two distinct eigenvalues, the Jordan form of  $A$  is diagonal, i.e.,  $J = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ . It is easy to see that the eigenvector  $u_1 = (x, y)^T$  associated with the eigenvalue  $\lambda_1 = 1$  satisfies the matrix equation

$$0 = (A - E)u_1 = \begin{pmatrix} -2 & 1 \\ -6 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

This given the equation  $-2x + y = 0$ , thus the eigenvectors are multiples of the vector  $u_1 = (1, 2)^T$ . Similarly, the eigenvector associated with the eigenvalue  $\lambda_2$  is given by  $u_2 = (1, 3)^T$ . The matrix  $P$  is then obtained by writing these eigenvectors into the columns,  $P = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$ . Then, for the matrix  $A$  we have  $A = PJP^{-1}$ . The inverse of  $P$  is  $P^{-1} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}$ , and we can confirm the relation  $A = PJP^{-1}$ :

$$\begin{pmatrix} -1 & 1 \\ -6 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}.$$

This decomposition says that the matrix  $A$  determines a linear mapping which has the aforementioned diagonal form with respect to the basis  $\{u_1, u_2\}$ . Geometrically, this means that in the direction of  $u_1$  nothing is changing and in the direction of  $u_2$  every vector is being stretched twice.

Let us now focus on  $B$ . This has a unique eigenvalue given by  $\lambda = 1$  with multiplicity 2. The corresponding eigenvector  $v = (x, y)^T$  satisfies the matrix equation

$$0 = (B - E)v = \begin{pmatrix} -2 & 1 \\ -4 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

We find that the solutions are multiples of the vector  $v = (1, 2)^T$ . The fact that the system does not have two linearly independent vectors as a solution, implies the following expression for the Jordan canonical form:  $J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ . The basis for which  $A$  has this form consists of the eigenvector  $v_1$  and the vector that is assigned to  $v_1$  via the linear transformation  $B - E$ . This provides a solution of the following system of equations:

$$\left( \begin{array}{cc|c} -2 & 1 & 1 \\ -4 & 2 & 2 \end{array} \right) \sim \left( \begin{array}{cc|c} -2 & 1 & 1 \\ 0 & 0 & 0 \end{array} \right).$$

You can easily check that the solutions are multiples of the vector  $v_2 = (1, 3)^T$ . Moreover, one obtains the same basis as in the previous case, and we can verify the relation  $B = PJP^{-1}$ :

$$\begin{pmatrix} -1 & 1 \\ -4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}.$$

In this case, the mapping acts on the vectors as follows: the component in the direction of  $v_2$  stays the same. On the other hand, the component in the direction of  $v_1$  is multiplied by the sum of the coefficients that determine the components in the directions of  $v_2$  and  $v_1$ . □

**3.F.59.** Consider the matrix  $A = \begin{pmatrix} -1 & -1 & -1 & 0 \\ 3 & 2 & 3 & -1 \\ 2 & 1 & 3 & -1 \\ 2 & 1 & 4 & -2 \end{pmatrix}$ . Solve the following tasks:

- i) Show that the eigenvalues of  $A$  are given by  $\pm 1$ . Compute their algebraic and geometric multiplicities and find the corresponding eigenvectors of  $A$ .
- ii) Use Sage to verify that the dimensions of the null spaces  $\ker((A - E)^j)$  do not increase for  $j > 3$ . Next verify that  $\dim \mathcal{R}_{\lambda_1} = 3$  by computing the rank of the matrix  $(A - E)^3$  in Sage and applying the rank-nullity theorem.

- iii) Derive the generalized eigenvectors of  $A$ .  
 iv) Describe the Jordan canonical form  $J$  of  $A$  and find a matrix  $P$  satisfying  $P^{-1}AP = J$ . Next verify your answer by Sage.  
 v) Prove the direct sum decomposition  $\mathbb{R}^4 = \mathcal{R}_{\lambda_1} \oplus \mathcal{R}_{\lambda_2}$ , where  $\lambda_1 = 1$  and  $\lambda_2 = -1$  are the eigenvalues of  $A$ .

**Solution.** (i) The characteristic polynomial of  $A$  is given by

$$\chi_A(\lambda) = \begin{vmatrix} -1 - \lambda & -1 & -1 & 0 \\ 3 & 2 - \lambda & 3 & -1 \\ 2 & 1 & 3 - \lambda & -1 \\ 2 & 1 & 4 & -2 - \lambda \end{vmatrix} = (\lambda - 1)^3(\lambda + 1).$$

To confirm this expression we may use Sage via the block

```
A=matrix([[ -1, -1, -1, 0], [3, 2, 3, -1], [2, 1, 3, -1], [2, 1, 4, -2]])
p=A.characteristic_polynomial(); show(factor(p))
```

Thus the eigenvalues of  $A$  are  $\lambda_1 = 1$  with multiplicity 3 and  $\lambda_2 = -1$  with multiplicity 1.

• Let us first focus on  $\lambda_1$  and describe the eigenspace  $V_{\lambda_1} = V_1 = \ker(A - E)$ . We will perform elementary row operations to obtain the reduced row echelon form of the matrix  $A - E$ :

$$\begin{aligned} A - E &= \begin{pmatrix} -2 & -1 & -1 & 0 \\ 3 & 1 & 3 & -1 \\ 2 & 1 & 2 & -1 \\ 2 & 1 & 4 & -3 \end{pmatrix} \xrightarrow{R_4 \rightarrow R_4 + R_1} \begin{pmatrix} -2 & -1 & -1 & 0 \\ 3 & 1 & 3 & -1 \\ 2 & 1 & 2 & -1 \\ 0 & 0 & 3 & -3 \end{pmatrix} \xrightarrow{R_3 \rightarrow R_3 + R_1} \begin{pmatrix} -2 & -1 & -1 & 0 \\ 3 & 1 & 3 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 3 & -3 \end{pmatrix} \\ &\xrightarrow{R_4 \rightarrow R_4 - 3R_3} \begin{pmatrix} -2 & -1 & -1 & 0 \\ 3 & 1 & 3 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \rightarrow -\frac{1}{2}R_1} \begin{pmatrix} 1 & 1/2 & 1/2 & 0 \\ 3 & 1 & 3 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_2 \rightarrow R_2 - 3R_1} \\ &\begin{pmatrix} 1 & 1/2 & 1/2 & 0 \\ 0 & -1/2 & 3/2 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \rightarrow R_1 + R_2} \begin{pmatrix} 1 & 0 & 2 & -1 \\ 0 & -1/2 & 3/2 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_2 \rightarrow -2R_2} \begin{pmatrix} 1 & 0 & 2 & -1 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &\xrightarrow{R_2 \rightarrow R_2 + 3R_3} \begin{pmatrix} 1 & 0 & 2 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \rightarrow R_1 - 2R_3} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \end{aligned}$$

You can verify this expression in Sage by adding the following cel to the initial block:

```
E=identity_matrix(4); AEprref=(A-E).rref(); show(AEprref)
```

Thus, the solutions of the matrix equation  $(A - E)u_1 = 0$  for some vector  $u_1 = (x_1, x_2, x_3, x_4)^T$  correspond to solutions of the linear system  $\{x_1 + x_4 = 0, x_2 - x_3 = 0, x_3 - x_4 = 0\}$ , where  $x_4$  is a free parameter. This gives

$$V_1 = \ker(A - E) = \{t(-1, 1, 1, 1)^T : t \in \mathbb{R}\}$$

and hence the geometric multiplicity of  $\lambda_1 = 1$  equals 1, i.e.,  $\gamma(\lambda_1) = 1 = \dim V_1$ . Moreover, the vector  $u_1 = (-1, 1, 1, 1)^T$  is an eigenvector of  $A$  corresponding to  $\lambda_1 = 1$ .

• For the second eigenvalue  $\lambda_2 = -1$  we have algebraic multiplicity 1 and thus also the geometric multiplicity should be 1, i.e.,  $\gamma(\lambda_2) = \dim V_{-1} = \dim \ker(A + E) = 1$ . This also implies that  $\dim \mathcal{R}_{\lambda_2} = 1$ , in particular  $\mathcal{R}_{\lambda_2} \cong V_{-1}$  (the dimension of the null spaces  $\ker((A + E)^\ell)$  stabilizes for  $\ell > 1$ ). We will apply elementary row operations to derive the reduced row echelon form of the matrix  $A + E$ .

$$\begin{aligned} A + E &= \begin{pmatrix} 0 & -1 & -1 & 0 \\ 3 & 3 & 3 & -1 \\ 2 & 1 & 4 & -1 \\ 2 & 1 & 4 & -1 \end{pmatrix} \xrightarrow{R_4 \rightarrow R_4 - R_3} \begin{pmatrix} 0 & -1 & -1 & 0 \\ 3 & 3 & 3 & -1 \\ 2 & 1 & 4 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \leftrightarrow R_2} \begin{pmatrix} 3 & 3 & 3 & -1 \\ 0 & -1 & -1 & 0 \\ 2 & 1 & 4 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_1 \rightarrow \frac{1}{3}R_1} \\ &\begin{pmatrix} 1 & 1 & 1 & -1/3 \\ 0 & -1 & -1 & 0 \\ 2 & 1 & 4 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_3 \rightarrow R_3 - 2R_1} \begin{pmatrix} 1 & 1 & 1 & -1/3 \\ 0 & -1 & -1 & 0 \\ 0 & -1 & 2 & -1/3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \xrightarrow{R_2 \rightarrow -R_2} \begin{pmatrix} 1 & 1 & 1 & -1/3 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 2 & -1/3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

$$\begin{array}{l}
 R_3 \rightarrow R_3 + R_2 \quad \left( \begin{array}{cccc} 1 & 1 & 1 & -1/3 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 3 & -1/3 \\ 0 & 0 & 0 & 0 \end{array} \right) \xrightarrow{R_3 \rightarrow \frac{1}{3}R_3} \left( \begin{array}{cccc} 1 & 1 & 1 & -1/3 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1/9 \\ 0 & 0 & 0 & 0 \end{array} \right) \xrightarrow{R_1 \rightarrow R_1 - R_2} \left( \begin{array}{cccc} 1 & 0 & 0 & -1/3 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1/9 \\ 0 & 0 & 0 & 0 \end{array} \right) \\
 R_2 \rightarrow R_2 - R_3 \quad \left( \begin{array}{cccc} 1 & 0 & 0 & -1/3 \\ 0 & 1 & 0 & 1/9 \\ 0 & 0 & 1 & -1/9 \\ 0 & 0 & 0 & 0 \end{array} \right).
 \end{array}$$

Once more we can verify this expression in Sage quickly, by adding the cell

```
AEmrref=(A+E).rref(); show(AEmrref)
```

This results in the system  $\{x_1 - \frac{1}{3}x_4 = 0, x_2 + \frac{1}{9}x_4 = 0, x_3 - \frac{1}{9}x_4 = 0\}$ , where  $x_4$  is a free parameter. Thus

$$V_{\lambda_2} = V_{-1} = \ker(A + E) = \left\{ t \left( \frac{1}{3}, -\frac{1}{9}, \frac{1}{9}, 1 \right) : t \in \mathbb{R} \right\}.$$

Therefore the corresponding eigenvector is given (up to a scalar) by  $v_1 = (\frac{1}{3}, -\frac{1}{9}, \frac{1}{9}, 1)^T$ .

(ii) We will use Sage to study the matrices  $(A-E)^2$ ,  $(A-E)^3$ , and so on. We proceed with the following block:

```
A=matrix([[ -1, -1, -1, 0], [3, 2, 3, -1], [2, 1, 3, -1], [2, 1, 4, -2]])
E=identity_matrix(4)
bb=(A-E)*(A-E)
show("\nThe matrix (A-E)^2 is given by:",bb)
cc=bb*(A-E)
show("\nThe matrix (A-E)^3 is given by:",cc)
dd=cc*(A-E)
show("\nThe matrix (A-E)^4 is given by:",dd)
ee=dd*(A-E)
show("\nThe matrix (A-E)^5 is given by:",ee)
```

Executing this block, we obtain

$$\begin{array}{l}
 \text{The matrix } (A-E)^2 \text{ is given by: } \begin{pmatrix} -1 & 0 & -3 & 2 \\ 1 & 0 & 2 & -1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -3 & 4 \end{pmatrix} \\
 \text{The matrix } (A-E)^3 \text{ is given by: } \begin{pmatrix} 0 & 0 & 3 & -3 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 9 & -9 \end{pmatrix} \\
 \text{The matrix } (A-E)^4 \text{ is given by: } \begin{pmatrix} 0 & 0 & -6 & 6 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & -2 & 2 \\ 0 & 0 & -18 & 18 \end{pmatrix} \\
 \text{The matrix } (A-E)^5 \text{ is given by: } \begin{pmatrix} 0 & 0 & 12 & -12 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 4 & -4 \\ 0 & 0 & 36 & -36 \end{pmatrix}
 \end{array}$$

This shows that the matrices  $(A-E)^4$  and  $(A-E)^5$  are (even) multiples of the matrix  $(A-E)^3$ . Thus, as we expected, the dimensions of the null subspaces  $\ker((A-E)^j)$  will not increase for  $j > 3 = \alpha(\lambda_1)$ . To compute the rank of the matrix  $(A-E)^3$  we can continue typing in the previous block the following syntax:

```
show("The rank of (A-E)^3 is given by:", cc.rank())
```

Sage's output has the form:

```
The rank of (A-E)^3 is given by:1.
```

Thus  $\text{rank}((A-E)^3) = 1$  and  $\dim \ker((A-E)^3) = 4 - 1 = 3 = \dim \mathcal{R}_{\lambda_1}$ , as it should be, since  $\alpha(\lambda_1) = 3$ .

(iii) Let us now determine generalized eigenvectors  $u_2, u_3 \in \mathcal{R}_{\lambda_1}$  for the first eigenvalue, where the algebraic multiplicity exceeds the geometric multiplicity. To do this, start with the matrix equation  $(A-E)u_2 = u_1$ , i.e.,

$$\begin{pmatrix} -2 & -1 & -1 & 0 \\ 3 & 1 & 3 & -1 \\ 2 & 1 & 2 & -1 \\ 2 & 1 & 4 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

where  $u_2 = (x_1, x_2, x_3, x_4)^T$  is the unknown. We will use Sage to obtain the reduced row echelon form of the augmented matrix  $(A - E \mid v_1)$ . It is sufficient to continue typing in our previous block and add the following cell

```
v1=vector([-1, 1, 1, 1])
B=(A-E).augment(v1, subdivide="True")
show(B.rref())
```

We see that

$$\left( \begin{array}{cccc|c} -2 & -1 & -1 & 0 & -1 \\ 3 & 1 & 3 & -1 & 1 \\ 2 & 1 & 2 & -1 & 1 \\ 2 & 1 & 4 & -3 & 1 \end{array} \right) \sim \left( \begin{array}{cccc|c} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right).$$

Now you may conclude that the solution space of the induced linear system has the form  $\{(x_1, x_2, x_3, x_4)^T = (-s, 1 + s, s, s)^T\}$ , where  $s$  is a free parameter. Therefore, without loss of generality we may assume that  $u_2 = (0, 1, 0, 0)^T$ . Next, using the matrix expression  $(A - E)^3$  provided above, we can easily verify that  $(A - E)^3 u_2 = 0$ .

Similarly, by solving the matrix equation  $(A - E)u_3 = u_2$  we find that  $u_3 = (1, -2, 0, 0)^T$  (up to a scalar). Verification of this result is left as an exercise for the reader. To summarize, the eigenvalue  $\lambda_1 = 1$  has the eigenvector  $u_1 = (-1, 1, 1, 1)^T$  and the generalized eigenvectors  $u_2 = (0, 1, 0, 0)^T$  and  $u_3 = (1, -2, 0, 0)^T$ .

(iv) It is easy to see that the generalized eigenvector  $u_3$  satisfies the additional relation  $(A - E)^2 u_3 = u_1$ . Hence the set

$$\{(A - E)^2 u_3, (A - E)u_3, u_3\} = \{u_1, u_2, u_3\}$$

is a Jordan chain of length 3 corresponding to  $\mathcal{R}_{\lambda_1}$ . Consequently, the Jordan form should contain a Jordan block of size 3 corresponding to  $\lambda_1$ ,  $J_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ , and a Jordan block of size 1 corresponding to  $\lambda_2$ ,  $J_{-1} = (-1)$ . This yields the following expression:

$$J = J_1 \oplus J_{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \oplus (-1) = \left( \begin{array}{ccc|c} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & -1 \end{array} \right).$$

Recall that in Sage a verification of the Jordan canonical form (up to order of the Jordan blocks) occurs by adding the cell

```
J= A.jordan_form(transformation=True); show(J)
```

Now, a similarity matrix  $P$  satisfying the condition  $P^{-1}AP = J$  has as columns the vectors  $\{u_1, u_2, u_3, v_1\}$  derived in (i) and (iii), i.e.,

$$P = \begin{pmatrix} -1 & 0 & 1 & 1/3 \\ 1 & 1 & -2 & -1/9 \\ 1 & 0 & 0 & 1/9 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

In particular,  $P$  is invertible since these vectors are linearly independent. While the computation of  $P^{-1}$  is left as an exercise, we use Sage to verify that  $P^{-1}AP = J$ :

```
J=matrix([[1, 1, 0, 0], [0, 1, 1, 0], [0, 0, 1, 0], [0, 0, 0, -1]])
P=matrix([[-1, 0, 1, 1/3], [1, 1, -2, -1/9], [1, 0, 0, 1/9], [1, 0, 0, 1]])
P.inverse()*A*P == J
```

Sage's output is True. Finally, recall from Chapter 2 that in Sage a quick way to verify the linear independence of the vectors  $u_1, u_2, u_3$  and  $v_1$  is by computing the determinant of  $P$ , via the command  $P.det()$ . Alternatively, you might prefer to use the cell provided below:



```

u1=vector([-1, 1, 1, 1])
u2=vector([0, 1, 0, 0])
u3=vector([1, -2, 0, 0])
v1=vector([1/3, -1/9, 1/9, 1])
V=RR^4; V.linear_dependence([u1, u2, u3, v1])==[]
    
```

(v) We have seen that  $\dim \mathcal{R}_1 = 3$  and  $\dim \mathcal{R}_{-1} = \dim V_{-1} = 1$ . Since the total dimension of  $\mathbb{R}^4$  equals 4, the generalized eigenspaces together have enough “space” to cover all of  $\mathbb{R}^4$ . In fact, above we saw that joining the bases  $\{u_1, u_2, u_3\}$  of  $\mathcal{R}_1$  and  $\{v_1\}$  of  $\mathcal{R}_{-1} \cong V_{-1}$  we obtain a basis of  $\mathbb{R}^4$ , and this already provides a proof. Alternatively, the relations  $\mathbb{R}^4 = \mathcal{R}_1 + \mathcal{R}_{-1}$  and  $\mathcal{R}_1 \cap \mathcal{R}_{-1} = \{0\}$  imply the direct sum decomposition

$$\mathbb{R}^4 = \mathcal{R}_{\lambda_1} \oplus \mathcal{R}_{\lambda_2} = \mathcal{R}_1 \oplus \mathcal{R}_{-1} = \ker((A - E)^3) \oplus \ker(A + E).$$

□

## E) MATERIAL ON LINEAR ALGEBRA – MATRIX DECOMPOSITIONS

**3.F.60.** Find the  $LU$ -decomposition of the matrix

$$A = \begin{pmatrix} 2 & 8 & 0 \\ 2 & 2 & -3 \\ 1 & 2 & 6 \end{pmatrix}.$$

**Solution.** An answer goes as follows:

$$A = \begin{pmatrix} 2 & 8 & 0 \\ 2 & 2 & -3 \\ 1 & 2 & 6 \end{pmatrix} = LU = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 8 & 0 \\ 0 & -6 & -3 \\ 0 & 0 & 7 \end{pmatrix}.$$

□

**3.F.61.** Provide the  $QR$ -decomposition of the following matrices:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

**Solution.** The  $QR$ -factorizations under question are given as follows:

$$A = QR = \begin{pmatrix} 1/2 & -\sqrt{12}/4 & 0 \\ 1/2 & \sqrt{12}/12 & -\sqrt{6}/3 \\ 1/2 & \sqrt{12}/12 & \sqrt{6}/6 \\ 1/2 & \sqrt{12}/12 & \sqrt{6}/6 \end{pmatrix} \begin{pmatrix} 2 & 3/2 & 1 \\ 0 & \sqrt{12}/4 & \sqrt{12}/6 \\ 0 & 0 & \sqrt{6}/3 \end{pmatrix},$$

$$B = QR = \begin{pmatrix} \sqrt{2}/2 & \sqrt{6}/6 & -\sqrt{3}/3 \\ \sqrt{2}/2 & -\sqrt{6}/6 & \sqrt{3}/3 \\ 0 & \sqrt{6}/3 & \sqrt{3}/3 \end{pmatrix} \begin{pmatrix} \sqrt{2} & \sqrt{2}/2 & \sqrt{2}/2 \\ 0 & \sqrt{6}/2 & \sqrt{6}/6 \\ 0 & 0 & 2\sqrt{3}/3 \end{pmatrix}.$$

□

**3.F.62. Ray-tracing.** In computer 3D-graphics the image is very often displayed using the Ray-tracing algorithm. The basis of this algorithm is an approximation of the light waves by a ray (line) and an approximation of the displayed objects by polyhedrons. These are bounded by planes and it is necessary to compute where exactly the light rays are reflected from these planes. From physics we know how the rays are reflected – the angle of impact equals the angle of reflection. We have already met this topic in the exercise ??.

The ray of light in the direction  $v = (1, 2, 3)$  hits the plane given by the equation  $x + y + z = 1$ . In what direction is it reflected?

**Solution.** The unit normal vector to the plane is  $n = \frac{1}{\sqrt{3}}(1, 1, 1)$ . The vector that gives the direction of the reflected ray  $v_R$  lies in the plane given by the vectors  $v, n$ . We can express it as a linear combination of these vectors. Furthermore, the rule for the angle of reflection says that  $\langle v, n \rangle = -\langle v_R, n \rangle$ . From there we obtain a quadratic equation for the coefficient of the linear combination.

This exercise can be solved in an easier, more geometric way. From the diagram we can derive directly that

$$v_R = v - 2\langle v, n \rangle n$$

. In our case,  $v_R = (-3, -2, -1)$ .

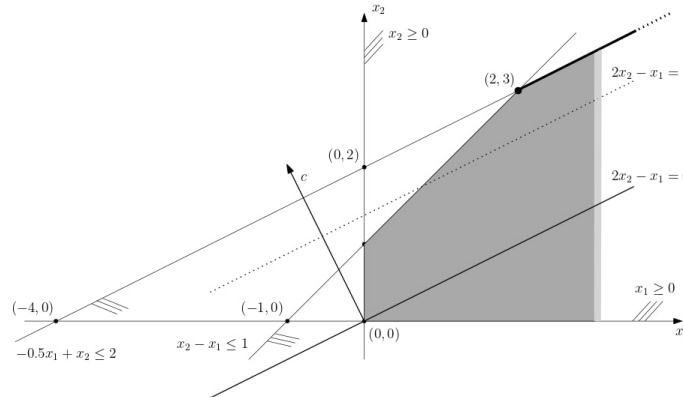
□

Solutions to the problems

**3.A.3.** We see that  $h = c^T x$ , where  $c = (-1, 2)^T$ . By multiplying the first inequality by  $-1$  we obtain the following equivalent system of constraints:

$$x_2 - x_1 \leq 1, \quad -0.5x_1 + x_2 \leq 2, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

An illustration of this LP problem goes as follows:



The feasible region (the grey region in this diagram) is unbounded. However, the drawn thick semi-line starting from the point  $(2, 3)$  give us infinitely many optimal solutions.

**3.A.5.** The daily diet should contain 3.9 kg of hay and 4.3 kg of oat. Then the costs per foal are €13.82.

**3.A.9.** (a) Multiply both sides of the second inequality by  $-1$  to get  $-x_1 - x_2 \leq 1$ . Also, replace the relation  $x_1 - 2x_2 = 1$  by the inequalities  $x_1 - 2x_1 \leq 1$  and  $x_1 - 2x_2 \geq 1$ . A multiplication of the latter by  $-1$  yields the standard form of the initial problem: maximize  $h = x_1 + 2.5x_2$  with respect to the conditions

$$2x_1 + 3x_2 \leq 20, \quad -x_1 - x_2 \leq 1, \quad x_1 - 2x_2 \leq 1, \quad -x_1 + 2x_2 \leq -1, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

Hence the dual problem is the minimization of  $20y_1 + y_2 + y_3 - y_4$  subject to the conditions

$$2y_1 - y_2 + y_3 - y_4 \geq 1, \quad 3y_1 - y_2 - 2y_3 + 2y_4 \geq 2.5, \quad y_i \geq 0, \quad \text{for all } i = 1, \dots, 4.$$

However, in the primal problem we had a constraint given by an equality, and so in the dual problem this should correspond to an unrestricted variable. Indeed, by setting  $y_5 = y_3 - y_4$  we result to the minimization of  $20y_1 + y_2 + y_5$  subject to

$$2y_1 - y_2 + y_5 \geq 1, \quad 3y_1 - y_2 - 2y_5 \geq 2.5, \quad y_1 \geq 0, \quad y_2 \geq 0,$$

with  $y_5$  unrestricted in sign.

(b) The standard form of the primal problem is the minimization of  $h = 2x_1 + 3x_2 + 2x_4$  subject to

$$-x_1 - 2x_2 - 2x_3 \geq 6, \quad -x_1 - 4x_2 + 2x_4 \geq -5, \quad x_1 + 4x_2 - 2x_4 \geq 5, \quad x_2 - x_3 + 4x_4 \geq 2, \quad x_i \geq 0, \quad \text{for all } i = 1, \dots, 4.$$

The dual problem reads as follows: maximize  $6y_1 - 5y_2 + 5y_3 + 2y_4$  subject to

$$-y_1 - y_2 + y_3 \leq 2, \quad -2y_1 - 4y_2 + 4y_3 + y_4 \leq 3, \quad -2y_1 - y_4 \leq 0, \quad 2y_2 - 2y_3 + 4y_4 \leq 2, \quad y_i \geq 0, \quad \text{for all } i = 1, \dots, 4.$$

By setting  $y_5 = y_2 - y_3$  we finally result to the maximization of  $6y_1 + 2y_4 - 5y_5$  with respect to

$$-y_1 - y_5 \leq 2, \quad -2y_1 + y_4 - 4y_5 \leq 3, \quad -2y_1 - y_4 \leq 0, \quad 4y_4 + 2y_5 \leq 2, \quad y_1 \geq 0, \quad y_4 \geq 0,$$

with  $y_5$  unrestricted in sign.

**3.B.4.** Let us proceed with Sage for now, although we recommend readers to follow the discussion in Section 3.2.2, and perform the formal computations alone. For the given recurrence relation, the corresponding code is as follows:

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n+2) - 3*a(n+1) - 3*a(n)
initial = {a(1):1, a(2):3}
rsolve(f, a(n), initial)
```

As an answer, Sage returns the following:

$$-\sqrt{21} * (3/2 - \sqrt{21}/2) * n/21 + \sqrt{21} * (3/2 + \sqrt{21}/2) * n/21.$$

In other words, the solution is given by

$$x_n = \frac{1}{\sqrt{21}} \left( \frac{3 + \sqrt{21}}{2} \right)^n - \frac{1}{\sqrt{21}} \left( \frac{3 - \sqrt{21}}{2} \right)^n.$$

**3.B.7.** A verification in Sage can be obtained as follows:

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function('a')
f = a(n+2) - 2*a(n+1) - 3*a(n)
initial = {a(0) : 0, a(1) : 1}
rsolve(f, a(n), initial)
```

Sage's output has the form  $-(-1) * n/4 + 3 * n/4$ .

**3.B.8.** To treat this task we begin with the characteristic equation, given by  $r^4 - r^3 - r + 1 = 0$ , or equivalently by  $(r - 1)^2(r^2 + r + 1) = 0$ . This has two complex roots given by

$$r_1 = -\frac{1}{2} + i\frac{\sqrt{3}}{2} = \cos(2\pi/3) + i\sin(2\pi/3), \quad r_2 = -\frac{1}{2} - i\frac{\sqrt{3}}{2} = \cos(2\pi/3) - i\sin(2\pi/3),$$

and the real root 1 which is double. Thus we can find basis of the solution space; this consists of the sequences

$$\{a_n\} := \left\{ \left( -\frac{1}{2} + i\frac{\sqrt{3}}{2} \right)^n \right\}_{n=1}^{\infty}, \quad \{b_n\} := \left\{ \left( -\frac{1}{2} - i\frac{\sqrt{3}}{2} \right)^n \right\}_{n=1}^{\infty},$$

together with  $\{n\}_{n=1}^{\infty}$  and the constant sequence  $\{1\}_{n=1}^{\infty}$ .

Our focus now shifts to establishing a real basis for the solution space, by replacing the two complex generators from the previous basis with sequences that are entirely real. These generators are “power series” whose elements are complex conjugates. We will discuss power series in Chapter 5. Now, it suffices to take as generators appropriate linear combinations of  $a_n, b_n$ , as in the Problem 3.B.5. This yields the following real basis:  $\{1\}_{n=1}^{\infty}, \{n\}_{n=1}^{\infty}, \{\cos(2n\pi/3)\}_{n=1}^{\infty}, \{\sin(2n\pi/3)\}_{n=1}^{\infty}$ .

**3.B.9.** The answer is given by the sequence  $(x_n)$  with general term  $x_n = -3(-1)^n - 2\cos(n \cdot (2\pi/3)) - 2\sqrt{3}\sin(n \cdot (2\pi/3))$ .

**3.B.11.** The general solution of the homogeneous equation is of the form  $a(-1)^n + b2^n$ . A particular solution is the constant  $-1/2$ . Therefore, the general solution of the given non-homogeneous equation without initial conditions has the form

$$a(-1)^n + b2^n - \frac{1}{2}.$$

Substituting in the initial conditions we obtain the constants  $a = -5/6, b = 5/6$ . Thus, the desired solution is given by the sequence  $x_n = -\frac{5}{6}(-1)^n + \frac{5}{6}2^n - \frac{1}{2}$ . To confirm this result by Sage use the block

```
from sympy import Function, rsolve
from sympy.abc import n
a = Function("a")
f = a(n+2) - a(n+1) - 2*a(n) - 1
int = {a(1) : 2, a(2) : 2}
rsolve(f, a(n), int)
```

**3.C.5.** In this case we have  $b_1 = 0, b_2 = 2 = b_3, b_4 = 1$ , and thus the Leslie matrix has the form

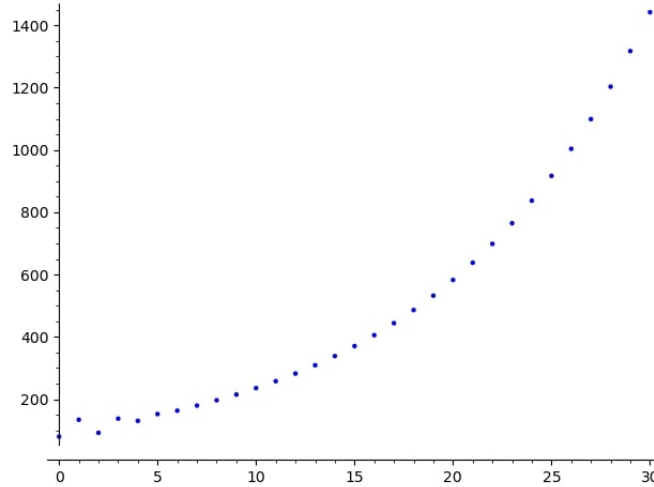
$$A = \begin{pmatrix} 0 & 2 & 2 & 1 \\ 0.4 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \end{pmatrix}.$$

Its unique dominant eigenvalue is  $\lambda_1 \approx 1.09486$ . Since  $\lambda_1 > 0$ , the population increases with growth rate  $1.09486^t$ , in contrast with the population of bushbabies in 3.C.3. The normalized eigenvector associated with  $\lambda_1$  is given by  $X_1 \approx (0.639, 0.233, 0.106, 0.019)^T$ , hence the long term trends of the female population are as follows: 63.9% of age class A, 23.3% of age class B, 10.6% of age class C and 1.9% of age class D. Finally, we compute

$$A^{10} \approx \begin{pmatrix} 1.047 & 2.872 & 2.094 & 0.96 \\ 0.384 & 1.047 & 0.761 & 0.348 \\ 0.174 & 0.48 & 0.349 & 0.16 \\ 0.032 & 0.087 & 0.064 & 0.029 \end{pmatrix}, \quad p_{10} = A^{10}p_0 \approx (150.813, 54.956, 25.171, 4.598)^T.$$

Therefore, the female population after ten years will consist of approximately 235 members (this is the sum of the entries of  $p_{10}$ ), and the total population of the colony (both females and males) after the same period will approximately reach the level

of 470 galagos. Below is the required diagram representing the exponential growth of the female population for a duration of thirty years.



**3.D.1.** Let us first recall that the norm of an arbitrary vector  $z \in \mathbb{C}^n$  with respect the standard Hermitian form  $\langle \cdot, \cdot \rangle$  reads by  $\|z\|^2 = \sum_{j=1}^n z_j \bar{z}_j$ . For the vectors  $x = (3 + 2i, 1 - i, -i)^T$  and  $y = (2 - 2i, 1 - i, 2 + i)^T$  in  $\mathbb{C}^3$  one computes

$$\langle x, y \rangle = y^* x = \begin{pmatrix} 2 + 2i & 1 + i & 2 - i \end{pmatrix} \begin{pmatrix} 3 + 2i \\ 1 - i \\ -i \end{pmatrix} = (3 + 2i)(2 + 2i) + (1 - i)(1 + i) - i(2 - i) = 3 + 8i.$$

Moreover, we see that

$$\begin{aligned} \|x\| &= \sqrt{(3 + 2i)(3 - 2i) + (1 - i)(1 + i) - i^2} = 4, \\ \|y\| &= \sqrt{(2 - 2i)(2 + 2i) + (1 - i)(1 + i) + (2 + i)(2 - i)} = \sqrt{15}. \end{aligned}$$

On the other hand, the vector  $x - y$  has the expression  $x - y = (1 + 4i, 0, -(2 + 2i))^T$ . Therefore,  $d(x, y) = \sqrt{(1 + 4i)(1 - 4i) + (2 + 2i)(2 - 2i)} = 5$ . Since  $\|x\| = 4$  and  $\|y\| = \sqrt{15}$ , the normalized vectors with  $\|\hat{x}\| = 1 = \|\hat{y}\|$  are given by  $\hat{x} = \frac{1}{4}(3 + 2i, 1 - i, -i)$  and  $\hat{y} = \frac{1}{\sqrt{15}}(2 - 2i, 1 - i, 2 + i)$ , respectively. See the Problem 3.D.2 how one can verify these results in Sage.

**3.D.3.** (a) For the vectors  $v = (1 + i, 2 - i)^T$  and  $w = (3 - 2i, 1 + i)^T$  in  $\mathbb{C}^2$  we compute

$$\langle v, w \rangle = w^* v = \begin{pmatrix} 3 + 2i & 1 - i \end{pmatrix} \begin{pmatrix} 1 + i \\ 2 - i \end{pmatrix} = (3 + 2i)(1 + i) + (1 - i)(2 - i) = 2 + 2i.$$

Moreover, we see that

$$\|v\|^2 = v^* v = (1 + i)(1 - i) + (2 - i)(2 + i) = 7, \quad \|w\|^2 = w^* w = (3 - 2i)(3 + 2i) + (1 + i)(1 - i) = 15.$$

Recall that the angle  $\theta$  between  $v, w$  is defined by the equation (compare with the formula given in 2.3.22 for the real case)

$$\cos \theta = \frac{\operatorname{Re}(\langle v, w \rangle)}{\|v\| \|w\|},$$

where  $\operatorname{Re}(z)$  denotes the real part of a complex number  $z \in \mathbb{C}$ . Hence it follows that  $\theta = \cos^{-1} \left( \frac{2}{\sqrt{7}\sqrt{15}} \right) \approx 1.37435 \approx 78.75^\circ$ . To verify all these results in Sage one can use the following cell:

```
v = vector(CDF, [1+I, 2-I])
w = vector(CDF, [3-2*I, 1+I])
vinnerw=w.hermitian_inner_product(v)
print("The standard Hermitian product of the vectors v=",v,"
and w=",w," "equals to",vinnerw)
vn=N(v.norm()^2, digits=3)
print("The square norm of the vector v=", v, "" "equals to", vn)
wn=N(w.norm()^2, digits=3)
print("The square norm of the vector w=", w, "" "equals to", wn)
theta=arccos(w.hermitian_inner_product(v).real()/(norm(v)*norm (w)))
```

```

print("The angle between the vectors v and w equals to",
N(theta, digits = 4), "radians" )
print("The angle between the vectors v and w equals to",
N(theta*180/pi, digits = 4), "degrees")
    
```

In this block we used the option CDF (Complex Double Field) when introducing the vectors  $v, w$ , similarly with the description given in 3.D.2. Recall that CDF approximates the field of complex numbers using double-precision floating point numbers. Alternatively, we could use the option CC (for this task, the result would remain the same). The CDF option helps to save some space in the output, compared to CC. Finally, for your convenience, in this block Sage was programmed to provide more detailed responses, than usual, as one can see in the output below:

```

The standard Hermitian product of the vectors v= (1.0 + 1.0*I, 2.0 - 1.0*I)
and w= (3.0 - 2.0*I, 1.0 + 1.0*I) equals to 2.0 + 2.0*I
The square norm of the vector v= (1.0 + 1.0*I, 2.0 - 1.0*I) equals to 7.00
The square norm of the vector w= (3.0 - 2.0*I, 1.0 + 1.0*I) equals to 15.0
The angle between the vectors v and w equals to 1.374 radians
The angle between the vectors v and w equals to 78.75 degrees
    
```

Similarly can be treated the cases (b) and (c) for which we only present the results.

(b)  $\langle v, w \rangle = 5 + 4i$ ,  $\|v\|^2 = 10$ ,  $\|w\|^2 = 7$ ,  $\cos \theta \approx 0.59$ ,  $\theta \approx 0.93 \approx 53.5^\circ$ .

(c)  $\langle v, w \rangle = 2 - 4i$ ,  $\|v\|^2 = 5$ ,  $\|w\|^2 = 19$ ,  $\cos \theta \approx 0.205$ ,  $\theta \approx 1.364 \approx 78.1^\circ$ .

**3.D.4.** One can easily check the following properties of  $f$ ;  $f(u + w, v) = f(u, v) + f(w, v)$ ,  $f(au, v) = af(u, v)$  and  $\overline{f(v, u)} = f(u, v)$ , for all  $u, v \in V = \mathbb{F}^2$  and  $a \in \mathbb{F}$ . Moreover, if  $u = (x_1, x_2)^T$  we have

$$f(u, u) = x_1\bar{x}_1 + 4x_1\bar{x}_2 + 4x_2\bar{x}_1 + x_2\bar{x}_2 = |x_1|^2 + 4(x_1\bar{x}_2 + x_2\bar{x}_1) + |x_2|^2,$$

where we used the relation  $z\bar{z} = |z|^2$  with  $z \in \mathbb{F}$ . However,  $x_1\bar{x}_2 + x_2\bar{x}_1 = |x_1 + x_2|^2 - |x_1|^2 - |x_2|^2$ , and hence the above relation reduces to  $f(u, u) = -3|x_1|^2 + 4|x_1 + x_2|^2 - 3|x_2|^2$ . Therefore, we may have  $f(u, u) < 0$  for some  $u \neq 0$ . For example,  $f(u, u) = -6$  for  $u = (1, -1)^T$ . This shows that  $f$  cannot be a scalar product on  $V$ .

**3.D.6.** We see that  $\rho_{a,b}(u, u) = au_1^2 + bu_2^2 \geq 0$  and the only way that this is zero is when  $u = 0$ , that is,  $u_1 = u_2 = 0$ . This proves positive definiteness. Moreover, by the commutativity of real numbers we get

$$\overline{\rho_{a,b}(v, u)} = \rho_{a,b}(v, u) = av_1u_1 + bv_2u_2 = \rho_{a,b}(u, v).$$

Hence  $\rho_{a,b}$  is Hermitian symmetric. Next,

$$\rho_{a,b}(u, v + w) = au_1(v_1 + w_1) + bu_2(v_2 + w_2) = (au_1v_1 + bu_2v_2) + (au_1w_2 + bu_2w_2) = \rho_{a,b}(u, v) + \rho_{a,b}(u, w),$$

and moreover  $\rho_{a,b}(cu, v) = \rho_{a,b}(u, cv) = c\rho_{a,b}(u, v)$  for any  $c \in \mathbb{R}$ . This proves that  $\rho_{a,b}$  is a scalar product on  $\mathbb{R}^2$ . Then the angle between  $u, v$  satisfies

$$\cos \theta = \frac{\rho_{a,b}(u, v)}{\|u\|_{\rho_{a,b}} \|v\|_{\rho_{a,b}}}.$$

For  $a = 2$  and  $b = 1$  we compute  $\rho_{2,1}(u, v) = 2 \cdot 1 + 1 \cdot (-1) = 1$ , and  $\|u\|_{\rho_{2,1}} = \sqrt{2 \cdot 1 + 1 \cdot 1} = \sqrt{3} = \|v\|_{\rho_{2,1}}$ . Thus  $\cos \theta = 1/3$  and so  $\theta = \arccos(1/3)$ . With respect to the dot product the vector  $u, v$  are orthogonal,  $u \cdot v = 0$ , hence they obviously form a different angle in comparison with the one appearing with respect to  $\rho_{2,1}$ .

**3.D.7.** For any vector  $x \in \mathbb{C}^n$  we see that  $\rho_A(x, x) = x^* A^* A x = (Ax)^* A x = \langle Ax, Ax \rangle \geq 0$ , where  $\langle \cdot, \cdot \rangle$  is the standard Hermitian form on  $\mathbb{C}^n$ . Thus  $\rho_A(x, x) \geq 0$  with the equality holding if and only if  $Ax = 0$ . Let us denote by  $A_1, \dots, A_n$  the  $n$  columns of  $A$ , such that  $A = (A_1 \ A_2 \ \dots \ A_n)$  with  $A_i = (a_{1i}, a_{2i}, \dots, a_{mi})^T$ , for any  $1 \leq i \leq n$ . Recall the column space of  $A$ :

$$C(A) = \{w \in \mathbb{C}^m : w = \sum_{i=1}^n x_i A_i, \text{ with } x_i \in \mathbb{C}\} = \{w \in \mathbb{C}^m : w = Ax\},$$

where  $x = (x_1, \dots, x_n)^T \in \mathbb{C}^n$ , and the second relation holds since the vector  $Ax \in \mathbb{C}^m$  is given by a linear combination of the columns of  $A$  with elements of  $x$ , i.e.,  $Ax = \sum_{i=1}^n x_i A_i$ . Hence our assumption that  $C(A)$  is  $n$ -dimensional, is equivalent to say that  $A$  consists of  $n$  linearly independent column vectors, or that the conditions  $Ax = 0$  and  $x = 0$  are equivalent each other. This shows that  $\rho_A$  is positive definite. Also, since  $\rho_A(x, y) \in \mathbb{C}$  for any  $x, y \in \mathbb{C}^n$  and so  $(\rho_A(x, y))^* = \overline{\rho_A(x, y)}$  we see that

$$\rho_A(x, y) = (Ay)^*(Ax) = ((Ax)^*(Ay))^* = (x^* A^* Ay)^* = (\rho_A(y, x))^* = \overline{\rho_A(y, x)},$$

for any  $x, y \in \mathbb{C}^n$ . Moreover,  $\rho_A(x, y + z) = (y + z)^*(A^* Ax) = y^* A^* Ax + z^* A^* Ax = \rho_A(x, y) + \rho_A(x, z)$  for any three vectors  $x, y, z$ , while for some  $a \in \mathbb{C}$  and any  $x, y \in \mathbb{C}^n$  we get  $\rho_A(ax, y) = (Ay)^* A ax = a(Ay)^* Ax = a\rho_A(x, y)$ . Thus  $\rho$  is an inner product.

We should mention that by the definition of  $\rho_A$  one gets  $\rho_A(x, y) = y^* A^* A x = (A y)^*(A x) = \langle A x, A y \rangle$ , where  $A x$  is viewed as a (column) vector. This provides of course an alternative and much shorter verification of the claim, but for the convenience of the reader we kept both arguments. Finally observe that when  $A$  is the identity matrix, then  $\rho_A$  is nothing than the standard Hermitian form on  $\mathbb{C}^n$ .

**3.D.9.** (a) For a real scalar product space  $(V, \langle \cdot, \cdot \rangle)$  we have  $\langle u, w \rangle = \langle w, u \rangle$ , for all  $u, w \in V$  and hence

$$\|u + w\|^2 = \langle u + w, u + w \rangle = \|u\|^2 + \|w\|^2 + 2\langle u, w \rangle, \quad u, w \in V.$$

The first claim is now direct.

For a complex scalar product space  $(V, \langle \cdot, \cdot \rangle)$  the bilinear form  $\langle \cdot, \cdot \rangle$  should satisfy  $\langle u, w \rangle = \overline{\langle w, u \rangle}$ , for all  $u, w \in V$ . This requirement prevents the form  $\langle \cdot, \cdot \rangle$  from having the same symmetry, as in the real case. For example, consider  $\mathbb{C}^2$  endowed with the standard (positive) Hermitian form  $\langle u, w \rangle = u_1 \bar{w}_1 + u_2 \bar{w}_2$ , with  $u, w \in \mathbb{C}^2$ . Given the vectors  $u = (0, i)^T \in \mathbb{C}^2$  and  $w = (0, 1)^T \in \mathbb{C}^2$  we see that  $\langle u, w \rangle = i$  and  $\|u + w\|^2 = 2$ ,  $\|u\|^2 = i\bar{i} = 1$ ,  $\|w\|^2 = 1$ . Hence we have  $\|u + w\|^2 = \|u\|^2 + \|w\|^2$  but  $u, w$  are not orthogonal (observe however that if  $\langle u, w \rangle = 0$ , then  $\|u + w\|^2 = \|u\|^2 + \|w\|^2$  still makes sense on a complex inner product space, see the proof of the theorem in 3.4.2).

(b) If  $u + w \perp u - w$  then  $0 = \langle u + w, u - w \rangle = \|u\|^2 - \|w\|^2$ . Thus  $\|u\|^2 = \|w\|^2$ , which gives  $\|u\| = \|w\|$ . Conversely, let  $u, w \in V$  such that  $\|u\| = \|w\|$ . Then  $\langle u + w, u - w \rangle = \|u\|^2 - \|w\|^2 = 0$ , hence  $u + w \perp u - w$ .

**3.D.10.** (a) From the definition of a unitary space given in 3.4.1, we see that

$$\|u + av\|^2 = \langle u + av, u + av \rangle = \langle u, u \rangle + \bar{a}\langle u, v \rangle + a\langle v, u \rangle + a\bar{a}\langle v, v \rangle = \|u\|^2 + 2\operatorname{Re}(\bar{a}\langle u, v \rangle) + |a|^2\|v\|^2$$

and similarly  $\|u - av\|^2 = \|u\|^2 - 2\operatorname{Re}(\bar{a}\langle u, v \rangle) + |a|^2\|v\|^2$ , for any two vector  $u, v \in V$  and scalar  $a$ . Thus,  $\|u + av\| = \|u - av\|$  if and only if  $\operatorname{Re}(\bar{a}\langle u, v \rangle) = 0$ . Since the latter relation should hold for any scalar  $a$ , taking  $a = 1$  for the real case, and  $a = i$  for the complex case, we see that  $\langle u, v \rangle = 0$ .

(b) The proof is similar and we leave it for practice.

**3.D.13.** According to part (4) of the main theorem in Section 3.4.2, for any two vectors  $x, y \in V$  we have the (Fourier) expansions  $x = \sum_j \langle x, E_j \rangle E_j$  and  $y = \sum_k \langle y, E_k \rangle E_k$ . Thus

$$\langle x, y \rangle = \left\langle \sum_j \langle x, E_j \rangle E_j, \sum_k \langle y, E_k \rangle E_k \right\rangle = \sum_{j,k} \langle x, E_j \rangle \overline{\langle y, E_k \rangle} \langle E_j, E_k \rangle = \sum_j \langle x, E_j \rangle \overline{\langle y, E_j \rangle}.$$

**3.D.14.** We can easily check that the given set  $\{E_1 = (-1, 1, 2), E_2 = (2, 0, 1), E_3 = (1, 5, -2)\}$  is an orthogonal basis of  $\mathbb{R}^3$ . We only prove orthogonality by Sage and leave a formal verification of the claim to the reader. Let us program Sage to do this quickly, as follows:

```
E1 = vector(QQbar, [-1, 1, 2])
E2 = vector(QQbar, [2, 0, 1])
E3 = vector(QQbar, [1, 5, -2])
B = [E1, E2, E3]
ips = [B[i].dot_product(B[j])
for i in range(2) for j in range(i+1, 2)]
all([ip == 0 for ip in ips])
```

Sage's answer is True. Notice in the same way we can also verify that the vectors  $E_i$  are *not* of unit length for all  $i = 1, 2, 3$ . This can be done by adding the cell

```
ips2 = [B[i].dot_product(B[i])
for i in range(2)]
all([ip2 == 1 for ip2 in ips2])
```

and in this case Sage prints out False.

Now recall that if  $\{u_1, \dots, u_m\}$  is an orthogonal basis of a unitary space  $(V, \langle \cdot, \cdot \rangle)$ , then any vector  $u \in V$  is expressed by

$$u = \sum_{k=1}^m \frac{\langle u, u_k \rangle}{\|u_k\|^2} u_k.$$

Let us apply this rule for our task. We compute

$$\langle u, E_1 \rangle = -12, \quad \langle u, E_2 \rangle = 8, \quad \langle u, E_3 \rangle = 24, \quad \|E_1\|^2 = 6, \quad \|E_2\|^2 = 5, \quad \|E_3\|^2 = 30.$$

Thus  $u = -\frac{12}{6}E_1 + \frac{8}{5}E_2 + \frac{24}{30}E_3$ .

**3.D.15.** We will apply a method that relies on the Cauchy-Schwarz inequality, see 3.4.3, as it is suggested in the statement. Set  $u = (\sqrt{a+2b}, \sqrt{b+2c}, \sqrt{c+2a})^T$ ,  $v = (1, 1, 1)^T$  and view  $u, v$  as vectors in  $\mathbb{R}^3$ . Their dot product is given by

$\langle u, v \rangle = v^T u = \sqrt{a+2b} + \sqrt{b+2c} + \sqrt{c+2a}$ , while  $\|u\|^2 = (a+2b) + (b+2c) + (c+2a) = 3(a+b+c)$  and  $\|v\|^2 = 3$ . An application of the Cauchy-Schwarz inequality  $\langle u, v \rangle \leq \|u\| \|v\|$  yields the expression

$$\sqrt{a+2b} + \sqrt{b+2c} + \sqrt{c+2a} \leq 3\sqrt{a+b+c},$$

which gives the result after dividing with  $\sqrt{a+b+c}$  (which is non-zero by the assumption in the statement).

**3.D.16.** The standard basis  $\underline{e}$  is an orthonormal basis of  $(\mathbb{C}^3, \langle \cdot, \cdot \rangle)$  and hence we see that

$$u = \sum_{j=1}^3 \langle u, e_j \rangle e_j = (2+i)e_1 + (-1+2i)e_2 + (3-i)e_3,$$

that is,  $\langle u, e_1 \rangle = 2+i$ ,  $\langle u, e_2 \rangle = -1+2i$  and  $\langle u, e_3 \rangle = 3-i$ . Now, according to Parseval's equality (see 3.4.3) we should have

$$\|u\|^2 = \sum_{j=1}^n |\langle u, e_j \rangle|^2 = |2+i|^2 + |-1+2i|^2 + |3-i|^2 = (2+i)(2-i) + (-1+2i)(-1-2i) + (3-i)(3+i) = 20,$$

where we applied the rule  $|z|^2 = z\bar{z}$ , with  $z \in \mathbb{C}$ . On the other hand, we also see that  $\|u\|^2 = \langle u, u \rangle = (2+i)\overline{(2+i)} + (-1+2i)\overline{(-1+2i)} + (3-i)\overline{(3-i)} = 20$ , and we are done.

**3.D.19.** When we rotate a point in  $\mathbb{R}^3$  about a particular axis, the coordinate corresponding to that axis remains unchanged. The remaining two coordinates are given by the well known rotation in the plane (see Chapter 1). This yields the matrices

$$R_x(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix}, \quad R_y(\theta) = \begin{pmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{pmatrix}, \quad R_z(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

which represent the rotation about the  $x$ -axis, the  $y$ -axis and the  $z$ -axis, respectively.

Note the sign of  $\theta$  in the rotation matrix about the  $y$ -axis. As with any other rotation, we want the rotation about the  $y$ -axis to be in the "positive sense". This means that when viewed from the negative  $y$ -axis direction, the rotation should appear anti-clockwise. Consequently, the signs in the rotation matrices depend on the orientation of our coordinate system.

Usually, in the 3-dimensional space the "right-handed coordinate system" is chosen, also known as "dextrorotary coordinate system". In a right-handed coordinate system, if you orient your right hand so that your thumb points along the positive  $x$ -axis and curl your fingers naturally, your fingers will point in the direction of the  $y$ -axis, followed by the  $z$ -axis. In particular, the index finger will point the positive  $y$ -axis and the middle finger will point the positive  $z$ -axis. For example, a positive rotation  $R_x(\theta)$  about the  $x$ -axis occurs perpendicular to the  $yz$ -plane, and a negative rotation will occur in the  $yz$ -plane, with opposite direction to the positive rotation. Notice when looking from positive  $x$  towards the origin, a positive rotation is counterclockwise. This configuration visually confirms the order  $x \rightarrow y \rightarrow z$  and its cyclical nature is evident in how axes are sequentially rotated:  $x \rightarrow y \rightarrow z \rightarrow x \rightarrow \dots$  for a positive cycle and  $x \rightarrow z \rightarrow y \rightarrow x \rightarrow \dots$  for a negative one.

**3.D.20.** The first task has been analyzed for unitary transformations in 3.4.4. Let us present an alternative proof. Suppose that  $U \in \text{Mat}_n(\mathbb{C})$  is unitary and consider the scalar product  $\langle Ux, y \rangle$ . Based on the relation  $E = UU^{-1} = U^{-1}U$ , where  $E$  is the identity  $n \times n$  matrix, we get  $\langle Ux, y \rangle = \langle Ux, UU^{-1}y \rangle = \langle x, U^{-1}y \rangle$ . On the other hand, by the definition of the conjugate transpose matrix we have  $\langle Ux, y \rangle = \langle x, U^*y \rangle$  for any  $x, y \in \mathbb{C}^n$ . By comparing these two relations and using bilinearity, we arrive at the following relation:

$$\langle x, (U^{-1} - U^*)y \rangle = 0, \quad (*)$$

for any  $x, y \in \mathbb{C}^n$ . We can now utilize the fact that  $\langle \cdot, \cdot \rangle$  is a non-degenerate bilinear form. Hence, by the relation (\*) we deduce that  $U^{-1} - U^* = 0$ , which implies that  $U^{-1} = U^*$ .

Conversely, assume that  $U^*U = E$ . Then, for any  $x \in \mathbb{C}^n$  we have  $\|Ux\|^2 = \langle Ux, Ux \rangle = \langle x, U^*Ux \rangle = \langle x, x \rangle = \|x\|^2$ , and hence  $U$  is unitary.

To prove that  $U(n)$  is a group, we can proceed similarly to the case of  $O(n)$ . Hence we need to verify that the composition of two unitary transformations and the inverse of a unitary transformation, both retain the same property. Let  $A, B \in U(n)$ . Then  $A^{-1} = A^*$  and  $B^{-1} = B^*$ , respectively. Moreover  $(AB)^* = B^*A^* = B^{-1}A^{-1} = (AB)^{-1}$ , and hence  $AB$  is also unitary. The verification of the second property is left as an exercise for the reader.

Finally, the claim for the determinant occurs by the relation  $\det(E) = 1 = \det(U^*U) = \det(U^*)\det(U) = \overline{\det(U)}\det(U) = |\det(U)|^2$ . This implies that the absolute value of the determinant of any unitary matrix is equal to one.

**3.D.22.** Consider linear maps  $\varphi, \psi : V \rightarrow V$  as in the statement, and let  $A, B$  be their matrices. In Problem 2.E.56 we proved that  $A, B$  must satisfy all the analogous properties listed in the statement. Since  $\varphi, \psi$  are determined by their matrices, all the formulas follow. Alternatively, one can work in terms of endomorphisms and prove all the relations based on the definition of the adjoint map. We leave such a description to the reader.



**3.D.25.** For  $A$  we have  $A^* = A$  and hence  $A$  is Hermitian (i.e., self-adjoint). We can directly verify this by Sage, as follows:

```
A=matrix(CC, [[sqrt(2), I, 1-I], [-I, 10, sqrt(2)+I], [1+I, sqrt(2)-I, 0]])
A.is_hermitian()
```

This command returns True. Alternatively, we can check this by verifying if  $A$  equals its conjugate transpose:

```
A=matrix(CC, [[sqrt(2), I, 1-I], [-I, 10, sqrt(2)+I], [1+I, sqrt(2)-I, 0]])
A == A.conjugate_transpose()
```

which again returns True.

Among the remaining matrices,  $B$  and  $C$  are Hermitian, while  $D$  is not. For the matrix  $B$ , verifying this property using Sage may take more time compared to a manual computation, but it is useful for handling parametric matrices. The verification process is as follows:

```
a = var("a"); b = var("b")
assume(a, "real"); assume(b, "real")
B = matrix(SR, 2, 2, [[a, I+b], [-I+b, sqrt(abs(a))]])
B.is_hermitian()
```

We leave the verification for the matrices  $C$  and  $D$  to the reader.

**3.D.26.** Under our assumptions for the matrices  $A, B$ , we need to show  $\text{tr}(A^*A) = \text{tr}(B^*B)$ . This is equivalent to proving that  $\sum_{i,j=1}^n |a_{ij}|^2 = \sum_{i,j=1}^n |b_{ij}|^2$ , providing an alternative reformulation of the problem. Since  $B = U^*AU$ , we have  $B^* = (U^*AU)^* = U^*A^*U$ . Therefore

$$\text{tr}(B^*B) = \text{tr}(U^*A^*UU^*AU) = \text{tr}(U^*A^*AU) = \text{tr}(A^*A).$$

This follows from the property that similar matrices  $X, Y$ , i.e.,  $X = P^{-1}YP$ , have the same trace,  $\text{tr}(X) = \text{tr}(Y)$ , and from the fact that  $U$  is unitary, i.e.,  $U^{-1} = U^*$ .

**3.D.27.** To compute its adjoint operator  $\varphi^* : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ , we need to find the matrix representation of  $\varphi^*$  with respect to the standard bases of  $\mathbb{R}^3$  and  $\mathbb{R}^2$ . In matrix form we have

$$\varphi \left( \begin{pmatrix} x \\ y \end{pmatrix} \right) = \begin{pmatrix} \sqrt{2} & 1 \\ 1 & -1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Thus, the matrix of  $\varphi$  with respect to the standard orthonormal bases is given by

$$A = \begin{pmatrix} \sqrt{2} & 1 \\ 1 & -1 \\ 0 & 2 \end{pmatrix}.$$

Now, to compute  $\varphi^*$  we can take the conjugate transpose of  $A$ , but since  $A$  has only real entries, this simplifies to just the transpose of  $A$ : We compute

$$A^* = A^T = \begin{pmatrix} \sqrt{2} & 1 & 0 \\ 1 & -1 & 2 \end{pmatrix}.$$

and hence  $\varphi^* \left( \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right) = \begin{pmatrix} \sqrt{2}x + y \\ x - y + 2z \end{pmatrix}$ , with  $x, y, z \in \mathbb{R}$ .

For the second task, we need to verify that  $B = B^*$ . The conjugate transpose of the matrix  $B$  is given by

$$B^* = \bar{B}^T = \begin{pmatrix} 10 & 5 + 5i & 3 + 2i \\ 5 - 5i & 5 & \sqrt{2}i \\ 3 - 2i & -\sqrt{2}i & 0 \end{pmatrix}.$$

Hence  $B = B^*$  and consequently  $\psi$  is a self-adjoint operator.

**3.D.28.** The adjoint operator of  $\text{tr} : \text{Mat}_m(\mathbb{F}) \rightarrow \mathbb{F}$  is a linear map  $\text{tr}^* : \mathbb{F} \rightarrow \text{Mat}_m(\mathbb{F})$ , assigning to any scalar  $z \in \mathbb{F}$  a  $m \times m$  matrix over  $\mathbb{F}$ . It is determined by the relation

$$\mathcal{B}(\text{tr}(A), z) = \mathcal{B}(A, \text{tr}^*(z)),$$

for any  $z \in \mathbb{F}$  and  $A \in \text{Mat}_m(\mathbb{F})$ . Observe here that on the left-hand side  $\mathcal{B}$  acts on scalars (viewed as trivial  $1 \times 1$  matrices). By the definition of  $\mathcal{B}$ , we have

$$\mathcal{B}(\text{tr}(A), z) = \bar{z} \text{tr}(A).$$

Setting  $\text{tr}^*(z) = B$  for some matrix  $B \in \text{Mat}_m(\mathbb{F})$ , we also obtain  $\mathcal{B}(A, \text{tr}^*(z)) = \text{tr}(B^*A)$ . Hence, we need  $\bar{z} \text{tr}(A) = \text{tr}(B^*A)$ . Based on the properties of the trace described in 2.C.38, we see that  $\bar{z} \text{tr}(A) = \text{tr}(\bar{z}A) = \text{tr}(\bar{z}EA)$ , where  $E$  denotes the  $m \times m$  matrix. By comparing these two relations and using the non-degeneracy of  $\mathcal{B}$ , we find that  $B^* = \bar{z}E$ . Therefore

$$B = (B^*)^* = (\bar{z}E)^* = E^* \bar{z} = Ez,$$

which means that  $\text{tr}^*(z) = zE$  from any  $z \in \mathbb{F}$ .

**3.D.29.** By definition, the linear map  $\text{proj}_W : V \rightarrow V$  sends  $u \in V$  to its orthogonal projection  $\text{proj}_W u$  on  $W$ . Let  $W^\perp$  be the orthogonal complement of  $W$  with respect to  $\langle \cdot, \cdot \rangle$ . Recall by Chapter 2 the direct sum decomposition  $V = W \oplus W^\perp$ , which for any  $u \in V$  implies that

$$u = \text{proj}_W u + \text{proj}_{W^\perp} u = \text{proj}_W u + (u - \text{proj}_W u) = P(u) + (u - P(u)),$$

where  $P(u) := \text{proj}_W u$ . Then, by linearity we see that

$$\langle P(u), w \rangle = \langle P(u), P(w) + (w - P(w)) \rangle = \langle P(u), P(w) \rangle + \langle P(u), w - P(w) \rangle,$$

for any  $u, w \in V$ . However, by assumption  $P(u) \in W$  and  $(w - P(w)) \in W^\perp$ . Thus, the second component in the right-hand side of the previous relation vanishes, which yields the relation  $\langle P(u), w \rangle = \langle P(u), P(w) \rangle$ . In a similar way one computes  $\langle u, P(w) \rangle = \langle P(u), P(w) \rangle$ , and the result follows.

**3.D.30.** The first claim is direct. To present a counterexample, use the matrix  $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \in \text{Mat}_3(\mathbb{R})$ . Then a direct computation shows that

$$A^* = A^T = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad AA^* = A^*A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Hence  $A$  is normal, but is not orthogonal since  $AA^T = A^T A \neq E$ .

**3.D.32.** We present directly the code including some comments within the code:

```
# Define the matrix phi symbolically
phi = Matrix(SR, [[2, 1+I, 0],
                 [1-I, 3, 0],
                 [0, 0, 1]])
# Verify that phi is normal
is_normal = phi*phi.conjugate_transpose() == phi.conjugate_transpose() * phi
print("Is phi normal?:", is_normal)
# Compute eigenvalues and eigenvectors symbolically
eigenvalues = phi.eigenvalues()
eigenvectors = phi.eigenvectors_right()
# Display eigenvalues and eigenvectors
print("Eigenvalues:", eigenvalues)
print("Eigenvectors:")
for eig in eigenvectors:
    print(f"Eigenvalue: {eig[0]}, Eigenvectors: {eig[1]}")
# Extract eigenvectors corresponding to distinct eigenvalues
if len(eigenvectors) >= 2:
    v1 = eigenvectors[0][1][0] # Eigenvector for the first eigenvalue
    v2 = eigenvectors[1][1][0] # Eigenvector for the second eigenvalue

    # Compute the inner product of the eigenvectors
    inner_product = v1.inner_product(v2)
    print("Inner product of v1 and v2:", inner_product)

    # Verify if they are orthogonal
    is_orthogonal = inner_product == 0
    print("Are v1 and v2 orthogonal?:", is_orthogonal)
else:
    print("Not enough distinct eigenvalues to check orthogonality.")
```

Let us also display Sage's output:

```
Is phi normal?: True
Eigenvalues: [1, 1, 4]
Eigenvectors:
Eigenvalue: 1, Eigenvectors: [(1, 1/2*I - 1/2, 0), (0, 0, 1)]
Eigenvalue: 4, Eigenvectors: [(1, -I + 1, 0)]
Inner product of v1 and v2: I + 1
Are v1 and v2 orthogonal?: (I + 1) == 0
```

Notice the comparison  $(I + 1) == 0$  is used to determine if the computed inner product is zero. Since  $1 + i \neq 0$  the result of the comparison is `False`, indicating that the eigenvectors are not orthogonal. We leave a formal computation for practice.

**3.D.34.** Such an example is given by the matrix  $A = \begin{pmatrix} 0 & 1-i \\ 1-i & 0 \end{pmatrix}$ . Indeed, since  $A^* = \bar{A}^T = \begin{pmatrix} 0 & 1+i \\ 1+i & 0 \end{pmatrix} \neq \pm A$ , the matrix at hand is neither Hermitian, nor skew-Hermitian. However,  $A$  is a normal matrix since

$$AA^* = \begin{pmatrix} 0 & 1-i \\ 1-i & 0 \end{pmatrix} \begin{pmatrix} 0 & 1+i \\ 1+i & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 1+i \\ 1+i & 0 \end{pmatrix} \begin{pmatrix} 0 & 1-i \\ 1-i & 0 \end{pmatrix} = A^*A.$$

Here is a detailed verification with Sage:

```
# Define the matrix
A = Matrix([[0, 1 - I], [1 - I, 0]])
# Compute the conjugate transpose (Hermitian transpose) of A
A_H = A.H
# Check if A is Hermitian
is_hermitian = A == A_H
# Check if A is skew-Hermitian
is_skew_hermitian = A == -A_H
# Compute A * A^H
A_AH = A * A_H
# Compute A^H * A
AH_A = A_H * A
# Check if A is normal
is_normal = A_AH == AH_A
# Print the results
print("Matrix A:")
print(A)
print("\nConjugate Transpose of A:")
print(A_H)
print("\nIs A Hermitian?")
print(is_hermitian)
print("\nIs A Skew-Hermitian?")
print(is_skew_hermitian)
print("\nIs A Normal?")
print(is_normal)
```

**3.D.35.** We will mainly use Sage and leave the formal computations for practice. The given matrix  $A$  is symmetric,  $A = A^T$ . The following cell verifies this statement in Sage:

```
A = matrix(SR, [[1, 2, 6], [2, 0, 2], [6, 2, 1]])
A.T == A
```

Recall also that the eigenvalues of  $A$  occur by the command

```
A.eigenvalues()
```

They are given by  $\lambda_1 = 8$ ,  $\lambda_2 = -5$  and  $\lambda_3 = -1$ . As for the eigenvectors, use the cell

```
A.eigenvectors_right()
```

Sage's output is the list

```
[(8, [(1, 1/2, 1)], 1), (-1, [(1, -4, 1)], 1), (-5, [(1, 0, -1)], 1)]
```

Thus the vectors  $u_1 = (2, 1, 2)^T$ ,  $u_2 = (1, 0, -1)^T$  and  $u_3 = (1, -4, 1)^T$  are eigenvectors of  $A$ , corresponding to the eigenvalues  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$ , respectively. Observe that  $\langle u_1, u_2 \rangle = 0$ ,  $\langle u_1, u_3 \rangle = 0$ , and  $\langle u_2, u_3 \rangle = 0$  where  $\langle \cdot, \cdot \rangle$  is the dot product on  $\mathbb{R}^3$ . Thus, the eigenvectors of  $A$  are orthogonal to each other. We also compute  $\|u_1\| = 3$ ,  $\|u_2\| = \sqrt{2}$  and  $\|u_3\| = 3\sqrt{2}$ . Consequently, the orthogonal matrix  $P$  has the form

$$P = \begin{pmatrix} \frac{2}{3} & \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ \frac{1}{3} & 0 & -\frac{4}{3\sqrt{2}} \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \end{pmatrix}.$$

Note that  $P$  is not uniquely defined. It remains to verify the relations  $P^{-1} = P^T$ , and  $P^{-1}AP = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ . In Sage this can be done easily:

```
A = matrix(SR, [[1, 2, 6], [2, 0, 2], [6, 2, 1]])
D = diagonal_matrix([8, -5, -1])
P = matrix(SR, [[2/3, 1/sqrt(2), 1/(3*sqrt(2))], [1/3, 0, -4/(3*sqrt(2))],
[2/3, -1/sqrt(2), 1/(3*sqrt(2))]])
print(P.inverse()==P.T); print(P.inverse()*A*P==D)
```

**3.D.36.** The one direction starting with  $A$  orthogonally diagonalizable is easy. Then we have  $P^TAP = D$  for some orthogonal  $n \times n$  matrix  $P$  and a diagonal  $n \times n$  matrix  $D$ . Thus  $A = PDP^T$  and since  $D$  being diagonal satisfies  $D = D^T$ , we obtain

$$A^T = (PDP^T)^T = PD^T P^T = PDP^T = A.$$

Hence  $A$  is symmetric. The converse is more difficult. Assume that  $A$  is a symmetric  $n \times n$  matrix. If  $n = 1$  the statement is trivial, i.e. the pair  $(P = 1, D = A)$  satisfies the statement. We will proceed by induction on the size of  $A$ . The induction hypothesis assumes that every  $(n-1) \times (n-1)$  symmetric matrix is orthogonally diagonalizable. Consider now the symmetric matrix  $A$  and let  $\lambda \in \mathbb{R}$  be an eigenvalue of  $A$  with eigenvector  $u \in \mathbb{R}^n$ , i.e.,  $Au = \lambda u$ . Set  $p_1 = u/\|u\|$  and extend  $\{p_1\}$  to an orthonormal basis  $\{p_1, \dots, p_n\}$  of  $\mathbb{R}^n$ . Let  $P$  be the orthogonal matrix whose column vectors are the vectors  $p_i$ , that is,  $P = (p_1 \ p_2 \ \dots \ p_n)$ . Then, in terms of block matrices for the multiplication  $AP$  we get the expression

$$AP = (\lambda p_1 \ Ap_2 \ \dots \ Ap_n) = P \begin{pmatrix} \lambda & C \\ 0 & A_1 \end{pmatrix},$$

for some  $C \in \text{Mat}_{1, n-1}(\mathbb{R})$  and  $A_1 \in \text{Mat}_{n-1}(\mathbb{R})$ . However,  $P^{-1} = P^T$  and hence this equivalently can be rewritten as

$$P^TAP = \begin{pmatrix} \lambda & C \\ 0 & A_1 \end{pmatrix}. \quad (*)$$

Since  $A$  is symmetric, we see that  $(P^TAP)^T = P^TAP$ , hence the matrix  $P^TAP$  is also symmetric. Thus, the relation given in  $(*)$  will make sense only when  $C = 0$  and  $A_1$  is symmetric. Since  $A_1$  has size  $(n-1) \times (n-1)$ , the induction hypothesis applies, so there exist an orthogonal matrix  $B$  and a diagonal matrix  $D_1$  of the same size with  $A_1$ , satisfying  $B^T A_1 B = D_1$ .

Consider now the block matrix  $P' = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix}$ , whose size is  $n \times n$ . Since  $B$  is orthogonal,  $P'$  is also orthogonal and hence we see that

$$(PP')^T A (PP') = (P')^T (P^TAP) P' = \begin{pmatrix} 1 & 0 \\ 0 & B^T \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & B^T A_1 B \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & D_1 \end{pmatrix}.$$

Since the last matrix is diagonal and  $PP'$  is orthogonal, the proof is complete.

**3.D.38.** Both matrices  $A$  and  $B$  are Hermitian, i.e.,  $A = A^*$  and  $B = B^*$ , and hence also normal. Thus, they are unitary diagonalizable. Moreover, since Hermitian matrices have only real eigenvalues, both  $A$  and  $B$  possess only real eigenvalues, as we will see also below. Let us begin with the matrix  $A$ . The characteristic polynomial of  $A$  is given by

$$\begin{aligned} \chi_A(\lambda) &= |A - \lambda E| = \begin{vmatrix} -(1+\lambda) & 0 & i \\ 0 & 1-\lambda & 0 \\ -i & 0 & 1-\lambda \end{vmatrix} = -(1+\lambda) \begin{vmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{vmatrix} + i \begin{vmatrix} 0 & 1-\lambda \\ -i & 0 \end{vmatrix} \\ &= -(1+\lambda)(1-\lambda)^2 + i^2(1-\lambda) = -(1+\lambda)(1-\lambda^2+1) = (1-\lambda)(\lambda^2-2) = (1-\lambda)(\lambda-\sqrt{2})(\lambda+\sqrt{2}). \end{aligned}$$

Or one may try to express the  $3 \times 3$  determinant with respect to the second row, which has less computations. Hence the eigenvalues of  $A$  are given by  $\lambda = 1$  and  $\lambda_{\pm} = \pm\sqrt{2}$ , all with multiplicity one. Hence their geometric multiplicity is also one. Let us derive generators of the corresponding eigenspaces  $V_{\lambda}$ ,  $V_{\lambda_+}$  and  $V_{\lambda_-}$ , which are subspaces of  $\mathbb{C}^3$ .

• For  $\lambda = 1$  we have the matrix equation  $Au = \lambda u = u$ , for some vector  $u = (u_1, u_2, u_3)^T \in \mathbb{C}^3$ , i.e.,

$$\begin{pmatrix} -u_1 + iu_3 \\ u_2 \\ -iu_1 + u_3 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}.$$

It is easy to see that the solutions of the corresponding linear system are given by  $u_1 = u_3 = 0$ , with  $u_2 \in \mathbb{C}$  free. Thus we may assume that  $u = (0, 1, 0)^T$ .

• For  $\lambda_+ = \sqrt{2}$  we need to solve the matrix equation  $Au_+ = \sqrt{2}u_+$ , for some  $u_+ = (w_1, w_2, w_3) \in \mathbb{C}^3$ , i.e.,

$$\begin{pmatrix} -w_1 + iw_3 \\ w_2 \\ -iw_1 + w_3 \end{pmatrix} = \begin{pmatrix} \sqrt{2}w_1 \\ \sqrt{2}w_2 \\ \sqrt{2}w_3 \end{pmatrix}.$$

This induces the linear system

$$\begin{cases} (\sqrt{2} + 1)w_1 - iw_3 = 0, \\ (\sqrt{2} - 1)w_2 = 0, \\ iw_1 + (\sqrt{2} - 1)w_3 = 0, \end{cases}$$

whose solutions are given by  $\{w_1 = i(\sqrt{2} - 1)r, w_2 = 0, w_3 = r \in \mathbb{C}\}$ . Thus, setting  $r = 1$  we may assume that  $u_+ = (i(\sqrt{2} - 1), 0, 1)^T \in \mathbb{C}^3$ .

• For  $\lambda_- = -\sqrt{2}$  we need to solve the matrix equation  $Au_- = -\sqrt{2}u_-$ , for some vector  $u_- = (v_1, v_2, v_3) \in \mathbb{C}^3$ . In a similar way we find that the solution of the corresponding linear system is given by  $\{v_1 = -i(\sqrt{2} + 1)s, v_2 = 0, v_3 = s \in \mathbb{C}\}$ . Thus, up to a multiple, the eigenvector  $u_-$  is given by  $u_- = (-i(\sqrt{2} + 1), 0, 1)^T \in \mathbb{C}^3$ .

We recommend formally verifying that the eigenvectors are orthogonal with respect to the standard Hermitian form  $\langle \cdot, \cdot \rangle$  on  $\mathbb{C}^3$ . We have achieved this using Sage, as presented below:

```
u=vector([0, 1, 0])
u_plus=vector([1-sqrt(2), 0, i])
u_minus=vector([1+sqrt(2), 0, i])
a=u.dot_product(u_plus.conjugate())
print(a.simplify_full())
b=u.dot_product(u_minus.conjugate())
print(b.simplify_full())
c=u_plus.dot_product(u_minus.conjugate())
print(c.simplify_full())
```

Now, without loss of generality, we can multiply both  $u_{\pm}$  by  $i$  and hence consider the eigenvectors

$$\tilde{u}_+ = \begin{pmatrix} 1 - \sqrt{2} \\ 0 \\ i \end{pmatrix} \in V_+, \quad \text{and} \quad \tilde{u}_- = \begin{pmatrix} 1 + \sqrt{2} \\ 0 \\ i \end{pmatrix} \in V_-.$$

We need to compute the norms of the eigenvectors  $u$ , and  $\tilde{u}_{\pm}$  with respect the standard Hermitian form  $\langle \cdot, \cdot \rangle$  on  $\mathbb{C}^3$ . Obviously,  $\|u\| = 1$  and

$$\begin{aligned} \|\tilde{u}_+\|^2 &= \langle \tilde{u}_+, \tilde{u}_+ \rangle = \tilde{u}_+^* \tilde{u}_+ = \begin{pmatrix} 1 - \sqrt{2} & 0 & -i \end{pmatrix} \cdot \begin{pmatrix} 1 - \sqrt{2} \\ 0 \\ i \end{pmatrix} = (1 - \sqrt{2})^2 - i^2 = 4 - 2\sqrt{2}, \\ \|\tilde{u}_-\|^2 &= \langle \tilde{u}_-, \tilde{u}_- \rangle = \tilde{u}_-^* \tilde{u}_- = \begin{pmatrix} 1 + \sqrt{2} & 0 & -i \end{pmatrix} \cdot \begin{pmatrix} 1 + \sqrt{2} \\ 0 \\ i \end{pmatrix} = (1 + \sqrt{2})^2 - i^2 = 4 + 2\sqrt{2}, \end{aligned}$$

where “ $\cdot$ ” denotes matrix multiplication. Thus  $\|\tilde{u}_+\| = \sqrt{4 - 2\sqrt{2}}$ ,  $\|\tilde{u}_-\| = \sqrt{4 + 2\sqrt{2}}$ , and the normalized eigenvectors are given by

$$\hat{u} = u = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \hat{u}_+ = \frac{\tilde{u}_+}{\|\tilde{u}_+\|} = \frac{1}{\sqrt{4 - 2\sqrt{2}}} \begin{pmatrix} 1 - \sqrt{2} \\ 0 \\ i \end{pmatrix}, \quad \hat{u}_- = \frac{\tilde{u}_-}{\|\tilde{u}_-\|} = \frac{1}{\sqrt{4 + 2\sqrt{2}}} \begin{pmatrix} 1 + \sqrt{2} \\ 0 \\ i \end{pmatrix}.$$

We can now derive the matrices  $U$ ,  $U^*$  (and  $D$ )

$$U = \begin{pmatrix} 0 & \frac{1 - \sqrt{2}}{\sqrt{4 - 2\sqrt{2}}} & \frac{1 + \sqrt{2}}{\sqrt{4 + 2\sqrt{2}}} \\ 1 & 0 & 0 \\ 0 & \frac{i}{\sqrt{4 - 2\sqrt{2}}} & \frac{i}{\sqrt{4 + 2\sqrt{2}}} \end{pmatrix}, \quad U^* = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1 - \sqrt{2}}{\sqrt{4 - 2\sqrt{2}}} & 0 & \frac{-i}{\sqrt{4 - 2\sqrt{2}}} \\ \frac{1 + \sqrt{2}}{\sqrt{4 + 2\sqrt{2}}} & 0 & \frac{-i}{\sqrt{4 + 2\sqrt{2}}} \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & -\sqrt{2} \end{pmatrix}.$$

Verifying the relation  $U^*AU = D$  by hand can be both time-consuming and challenging. To streamline this process we will use Sage in the following straightforward manner:

```
A=matrix([[ -1, 0, i], [0, 1, 0], [ -i, 0, 1]])
U=matrix([[0, (1-sqrt(2))/sqrt(4-2*sqrt(2)), (1+sqrt(2))/sqrt(4+2*sqrt(2))],
[1, 0, 0], [0, i/sqrt(4-2*sqrt(2)), i/sqrt(4+2*sqrt(2))]])
D=diagonal_matrix([1, sqrt(2), -sqrt(2)])
U.H*A*U==D
```

Verify that Sage's response is rapid, and effectively confirms the desired relation  $U^*AU = D$ .

For the matrix  $B$  we will rely solely on Sage to obtain a solution. Using the cell

```
B = matrix(QQbar, [[3, I], [-I, 3]])
B.eigenvalues()
B.eigenvectors_right()
```

we obtain the eigenvalues of  $B$ :  $\lambda_1 = 2$  and  $\lambda_2 = 4$ , both with multiplicity one. This code also returns the corresponding eigenvectors:  $(1, i)^T$  for  $\lambda_1$  and  $(1, -i)^T$  for  $\lambda_2$ . We may multiply the second vector by  $i$  and set  $u_1 = (1, i)^T$  and  $u_2 = (i, 1)^T$ . With respect to the standard Hermitian form on  $\mathbb{C}^2$  it is easy to verify that  $\langle u_1, u_2 \rangle = 0$ . Thus the eigenvectors are orthogonal, as it should be, and moreover  $\|u_1\| = \sqrt{2} = \|u_2\|$ . By taking the orthonormal vectors  $\hat{u}_1 = \frac{1}{\sqrt{2}}u_1$  and  $\hat{u}_2 = \frac{1}{\sqrt{2}}u_2$  we arrive to the unitary matrix  $U$  given by

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}.$$

Recall that Sage can verify this claim via the command `U.is_unitary`, where the same cell should include the definition of the matrix  $U$ . To verify that this matrix  $U$  diagonalizes the given matrix  $B$ , i.e.  $U^*BU = D$ , where  $D = \text{diag}(\lambda_1, \lambda_2) = \text{diag}(2, 4)$ , use the following block:

```
B = matrix(QQbar, [[3, I], [-I, 3]])
U = matrix(QQbar, [[1/sqrt(2), I*(1/sqrt(2))], [I*(1/sqrt(2)), 1/sqrt(2)]])
D = matrix(QQbar, [[2, 0], [0, 4]])
U.conjugate_transpose()*B*U == D
```

**3.D.39.** Suppose that  $A$  is a normal matrix with purely imaginary eigenvalues. Then we can find a unitary matrix  $U$  satisfying  $U^*AU = D$ , where  $D$  is a diagonal matrix with purely imaginary (diagonal) entries. From this we get  $A = UDU^*$  and since  $D^* = -D$ , we see that

$$A^* = (UDU^*)^* = UD^*U^* = -UDU^* = -A.$$

**3.D.41.** The matrices that represent Jordan canonical forms are  $G_1, G_2, G_5$  and  $G_6$ .

**3.D.43.** We saw that only the matrices  $G_1, G_2, G_5$  and  $G_6$  represent a Jordan canonical form. Let us denote by  $A_k$  the corresponding matrices and by  $P_k$  the corresponding similarity matrices. Both  $A_k$  and  $P_k$  have the same size as  $G_k$  for all these  $k$ . Since  $A_k = P_k G_k P_k^{-1}$ , i.e., the matrices  $A_k$  and  $G_k$  are similar for all  $k = 1, 2, 5, 6$ , they have the same characteristic polynomial. Thus  $\chi_{A_k}(\lambda) = \chi_{G_k}(\lambda)$  for all these four cases and from the expressions of  $G_k$  is direct that

- $\chi_{G_1}(\lambda) = \chi_{A_1}(\lambda) = \lambda^2(\lambda - 2)^2$ ;
- $\chi_{G_2}(\lambda) = \chi_{A_2}(\lambda) = (\lambda - 3)(\lambda - 1)^2$ ;
- $\chi_{G_5}(\lambda) = \chi_{A_5}(\lambda) = (\lambda - \pi)$ ;
- $\chi_{G_6}(\lambda) = \chi_{A_6}(\lambda) = (\lambda - 2)^2(\lambda - 1)^2\lambda^2$ .

**3.D.45.** (a) Let us compute the characteristic polynomial of the given matrix  $A = \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}$ :

$$\chi_A(\lambda) = |A - \lambda E| = \begin{vmatrix} 3 - \lambda & 1 \\ -1 & 1 - \lambda \end{vmatrix} = (3 - \lambda)(1 - \lambda) + 1 = \lambda^2 - 4\lambda + 4 = (\lambda - 2)^2.$$

Thus,  $\lambda = 2$  is the unique eigenvalue of  $A$ , and has algebraic multiplicity  $\alpha(\lambda) = 2$ . To determine the eigenvectors, we need to solve the matrix equation  $(A - 2E)u = 0$  for some vector  $u = (x, y)^T$ . This gives the linear system  $\{x + y = 0, -x - y = 0\}$  with 1-dimensional solution space  $\left\{ \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -s \\ s \end{pmatrix} : s \in \mathbb{R} \right\}$ . Therefore, we may assume that the eigenspace  $V_\lambda = V_2$  is generated by the eigenvector  $u = (1, -1)^T$  (this occurs for  $s = -1$ ). For the geometric multiplicity of  $\lambda$  this fact implies that  $\gamma(\lambda) = 1$ . Hence  $\gamma(\lambda) \neq \alpha(\lambda)$  and  $A$  cannot be diagonalizable. Since  $A$  is  $2 \times 2$ ,  $\alpha(\lambda) = 2$ ,  $\gamma(\lambda) = 1$ , and  $\lambda$  is the unique eigenvalue of  $A$ , the Jordan canonical form of  $A$  is the matrix  $J = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ .

Observe now that  $(A - 2E)^2 = 0$ , where 0 denotes the  $2 \times 2$  matrix. To find the generalized eigenvector corresponding to  $\lambda$  we need to solve the matrix equation  $(A - 2E)w = u$ . If we assume that  $w$  has the expression  $w = (w_1, w_2)^T$ , the solution space of the corresponding linear system has the form  $\left\{ \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1-t \\ t \end{pmatrix} : t \in \mathbb{R} \right\}$ . Thus we may assume that the generalized eigenvector  $w$  has the form  $w = (1, 0)^T$  (this occurs for  $t = 0$ ). Consequently, the basis with respect to which  $J$  has the above form is given by  $\{u, w\}$  and  $P$  has the form  $P = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$ . We also compute  $P^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$  and it is straightforward to verify the relation  $P^{-1}AP = J$ . In Sage this verification can be done as usual, i.e.,

```
A=matrix(SR, [[3, 1], [-1, 1]])
J=matrix(SR, [[2, 1], [0, 2]])
P=matrix(SR, [[1, 1], [-1, 0]])
Pinv=P.inverse(); Pinv*A*P==J
```

As mentioned in 3.D.44, an alternative has the form

```
A=matrix(SR, [[3, 1], [-1, 1]])
J, P = A.jordan_form(transformation=True)
show(J, P)
```

In this case Sage directly prints out the derived expression for  $J$  and  $P$ .

(b) For this task, observe that the  $k$ th power  $A^k$  of  $A$  satisfies the relation

$$A^k = \underbrace{(PJP^{-1})(PJP^{-1}) \cdots (PJP^{-1})}_{k\text{-factors}} = PJ^k P^{-1}.$$

Now, we easily compute  $J^4 = \begin{pmatrix} 16 & 32 \\ 0 & 16 \end{pmatrix}$  and by applying the relation  $A^4 = PJ^4 P^{-1}$ , we obtain  $A^4 = \begin{pmatrix} 48 & 32 \\ -32 & 16 \end{pmatrix}$ .

**3.D.46.** The matrix  $A$  is upper triangular, meaning that all entries below the main diagonal are zero. For an upper triangular matrix the characteristic polynomial is factorized into linear factors by inspecting the diagonal elements. For our case we get the expression  $\chi_A(\lambda) = (5 - \lambda)^4$ . Therefore,  $\lambda = 5$  is the unique eigenvalue of  $A$  and has algebraic multiplicity 4,  $\alpha(\lambda) = 4$ . To determine the size of the Jordan blocks, we need to find the geometric multiplicity of  $\lambda$ . The matrix  $A - 5E$  is given by

$$A - 5E = \begin{pmatrix} 0 & 4 & 2 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and it is easy to see that its rank equals 3. We can verify this statement in Sage, by the following block,

```
A = Matrix([[5, 4, 2, 1], [0, 5, 1, 1], [0, 0, 5, 2], [0, 0, 0, 5]])
E=identity_matrix(4)
AA=A-5*E; show(AA)
show(AA.rref())
print("The rank of A-4E is given by:", AA.rank())
```

Since  $\text{rank}(A - 4E) = 3$ , the dimension of  $\text{Ker}(A - 5E)$  equals 1.<sup>23</sup> Hence we conclude that the geometric multiplicity of  $\lambda = 5$  equals 1. Since  $\gamma(\lambda) = 1$ ,  $\alpha(\lambda) = 4$  and  $\lambda = 5$  is the unique eigenvalue of  $A$ , the Jordan form is composed of a single Jordan block of size 4 (see also below for a Jordan chain of length 4)

$$J = \begin{pmatrix} 5 & 1 & 0 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 5 \end{pmatrix}.$$

Let us now describe an eigenvector of  $\lambda$ . The previous block also provides the REEF of  $A - 4E$ , given by

$$A - 5E = \begin{pmatrix} 0 & 4 & 2 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

<sup>23</sup>Recall that according to the rank-nullity theorem we have  $\text{rank}(T) + \dim \text{Ker}(T) = \dim V$  for any linear transformation  $T : V \rightarrow W$  defined on a finite-dimensional vector space  $V$ .

Thus, the general solution of the linear system corresponding to the matrix equation  $(A - 4E)u_1 = 0$  is given by  $\{(x_1, x_2, x_3, x_4)^T = (t, 0, 0, 0)^T\}$ , where  $t$  is a free parameter. Hence, without loss of generality, we may assume that the eigenspace  $V_\lambda = V_5 = \text{Ker}(A - 5E)$  is generated by the eigenvector  $u_1 = (1, 0, 0, 0)^T$ . Using Sage we can confirm that  $(A - 5E)^4 = 0$  and compute

$$(A - 5E)^2 = \begin{pmatrix} 0 & 0 & 4 & 8 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (A - 5E)^3 = \begin{pmatrix} 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Therefore it is easy to derive the dimensions of the null spaces  $\text{Ker}((A - 5E)^2)$  and  $\text{Ker}((A - 5E)^3)$ :

$$\begin{aligned} \text{rank}((A - 5E)^2) = 2 &\implies \dim \text{Ker}((A - 5E)^2) = 4 - 2 = 2, \\ \text{rank}((A - 5E)^3) = 1 &\implies \dim \text{Ker}((A - 5E)^3) = 4 - 1 = 3. \end{aligned}$$

Since  $(A - 5E)^4 = 0$  it is immediate that  $\dim \text{Ker}((A - 5E)^4) = 4$  and hence  $\text{Ker}((A - 5E)^4)$  is the whole  $\mathbb{R}^4$ .

Let us compute a generalized eigenvector for the unique eigenvalue of  $A$ , which we will denote by  $u_2 = (x_1, x_2, x_3, x_4)^T$ . The matrix equation  $(A - 5E)u_2 = u_1$  corresponds to the linear system  $\{4x_2 + 2x_3 + x_4 = 1, x_3 + x_4 = 0, 2x_4 = 0\}$  and its solutions are given by  $\{(x_1, x_2, x_3, x_4)^T = (s, \frac{1}{4}, 0, 0)^T\}$ , where  $s$  is a free parameter. Thus, we may assume that  $u_2 = (0, \frac{1}{4}, 0, 0)^T$ , and it is easy to see that  $u_2 \in \text{Ker}((A - 5E)^2)$  but  $u_2 \notin \text{Ker}(A - 5E)$ . In particular, the vectors  $u_1, u_2$  span the subspace  $\text{Ker}((A - 5E)^2)$ . Next we need to solve the matrix equation  $(A - 5E)u_3 = u_2$  for some vector  $u_3 = (x_1, x_2, x_3, x_4)^T$ , which induces the system  $\{4x_2 + 2x_3 + x_4 = 0, x_3 + x_4 = 1/4, 2x_4 = 0\}$ . Its solutions are given by  $\{(x_1, x_2, x_3, x_4)^T = (r, -\frac{1}{8}, \frac{1}{4}, 0)^T\}$ , where  $r$  is a free parameter. Thus we may assume that  $u_3 = (0, -\frac{1}{8}, \frac{1}{4}, 0)^T$  and it is easy to see that  $u_3 \in \text{Ker}((A - 5E)^3)$  but  $u_3 \notin \text{Ker}((A - 5E)^2)$ . It follows that vectors  $u_1, u_2, u_3$  generate  $\text{Ker}((A - 5E)^3)$ . Finally, solving the matrix equation  $(A - 5E)u_4 = u_3$  for some unknown  $u_4 = (x_1, x_2, x_3, x_4)^T$ , we get the generalized eigenvector  $u_4 = (0, \frac{3}{32}, -\frac{1}{4}, \frac{1}{8})^T$ . Putting all information together, we deduce that  $\text{Ker}((A - 5E)^4) = \text{span}\{u_1, u_2, u_3, u_4\}$ .

To summarize, we saw that the vectors  $u_4, u_3, u_2$  and  $u_1$  satisfy the relations

$$\begin{array}{l} (A - 5E)^4 u_4 = 0 \mid (A - 5E)^3 u_3 = 0 \mid (A - 5E)^2 u_2 = 0 \mid (A - 5E)u_1 = 0, \\ (A - 5E)u_4 = u_3 \mid (A - 5E)u_3 = u_2 \mid (A - 5E)u_2 = u_1. \end{array}$$

Moreover, a direct computations shows that

$$(A - 5E)^3 u_4 = u_1, \quad (A - 5E)^2 u_4 = u_2, \quad (A - 5E)u_4 = u_3.$$

Thus  $\{(A - 5E)^3 u_4, (A - 5E)^2 u_4, (A - 5E)u_4, u_4\} = \{u_1, u_2, u_3, u_4\}$  is a Jordan chain of full length 4. This implies that the generalized eigenspace  $\mathcal{R}_\lambda$  corresponding to the eigenvalue  $\lambda$  has dimension 4 and it encompasses all vectors that map to zero under the fourth power of  $(A - 5E)$ . Thus,

$$\mathcal{R}_\lambda = \text{Ker}((A - 5E)^4) \cong \mathbb{R}^4.$$

Actually  $\mathcal{R}_\lambda$  is the union of the subspaces  $\text{Ker}((A - 5E)^\ell)$  for all  $1 \leq \ell \leq 4$ .

As another consequence of this chain we can derive a representative of the matrix  $P$ :

$$P = (u_1 \mid u_2 \mid u_3 \mid u_4) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/4 & -1/8 & 3/32 \\ 0 & 0 & 1/4 & -1/4 \\ 0 & 0 & 0 & 1/8 \end{pmatrix}.$$

Check yourselves using Sage via the following block that  $P$  satisfies  $P^{-1}AP = J$ .

```
A = matrix(SR, [[5, 4, 2, 1], [0, 5, 1, 1], [0, 0, 5, 2], [0, 0, 0, 5]])
J = matrix(SR, [[5, 1, 0, 0], [0, 5, 1, 0], [0, 0, 5, 1], [0, 0, 0, 5]])
P = matrix(SR, [[1, 0, 0, 0], [0, 1/4, -1/8, 3/32], [0, 0, 1/4, -1/4],
[0, 0, 0, 1/8]])
#Verify P^(-1)AP = J
print(P.inverse()*A*P == J)
```

**3.E.5.** (i) In 3.E.1 the matrix  $U$  was derived from the matrix  $A$  by successively applying the elementary row operations  $R_2 \rightarrow R_2 - 2R_1$ ,  $R_3 \rightarrow R_3 - 3R_1$  and  $R_3 \rightarrow R_3 + R_2$ . To obtain the corresponding elementary matrices, we apply each of these row operations to the  $3 \times 3$  identity matrix. This results in the following matrices:

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$



and the Gauss elimination in terms of matrices has the form

$$E_3 E_2 E_1 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -2 & 1 & 0 \\ -4 & 4 & 2 \\ -6 & 1 & -1 \end{pmatrix} = \begin{pmatrix} -2 & 1 & 0 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{pmatrix} = U.$$

Since the elementary matrices are always invertible, we obtain  $A = E_1^{-1} E_2^{-1} E_3^{-1} U$ , which implies that  $L = E_1^{-1} E_2^{-1} E_3^{-1}$ . The inverses of the elementary matrices are easily computed, leading to the expected result:

$$L = E_1^{-1} E_2^{-1} E_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -1 & 1 \end{pmatrix}.$$

We also compute

$$L^{-1} = E_3 E_2 E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -5 & 1 & 1 \end{pmatrix}.$$

For a verification using Sage, type the cell

```
L=matrix([[1, 0, 0], [2, 1, 0], [3, -1, 1]]); show(L.inverse())
```

(ii) To find the inverse of the matrix  $U$  via the Gauss elimination method, we begin by forming the augmented matrix  $(U | E)$ , where  $E$  is the  $3 \times 3$  identity matrix. We then successively apply the following elementary row operations:  $R_1 \rightarrow -\frac{1}{2}R_1$ ,  $R_1 \rightarrow R_1 + \frac{1}{4}R_2$ ,  $R_1 \rightarrow R_1 - \frac{1}{2}R_3$ ,  $R_2 \rightarrow \frac{1}{2}R_2$  and  $R_2 \rightarrow R_2 - R_3$ . You should be able to verify on your own that this results in the matrix

$$U^{-1} = \begin{pmatrix} -1/2 & 1/4 & -1/2 \\ 0 & 1/2 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here is a quick verification using Sage:

```
U=matrix([[-2, 1, 0], [0, 2, 2], [0, 0, 1]]); show(U.inverse())
```

(iii) We have  $\det(A) = \det(L) \det(U) = 1 \cdot (-4) = -4 \neq 0$ , and hence  $A$  is invertible. Now, using the expressions of  $L^{-1}$  and  $U^{-1}$  obtained above, the matrix  $A^{-1}$  occurs very easily:

$$A^{-1} = (LU)^{-1} = U^{-1}L^{-1} = \begin{pmatrix} -1/2 & 1/4 & -1/2 \\ 0 & 1/2 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -5 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 3/2 & -1/4 & -1/2 \\ 4 & -1/2 & -1 \\ -5 & 1 & 1 \end{pmatrix}.$$

Verify this result in Sage with the same basic method applied for  $L$  or  $U$ .

**3.E.6.** It is easy to see that  $\text{rank}(A) = 2$  and hence there exist two non-zero singular values. Since  $A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$ , the characteristic polynomial of  $A^T A$  is given by  $\chi_{A^T A}(\lambda) = (40 - \lambda)(10 - \lambda)$ . Hence we conclude that the eigenvalues of  $A^T A$  are  $\lambda_1 = 40$  and  $\lambda_2 = 10$ , so the singular values of  $A$  are  $\sigma_1 = \sqrt{40} = 2\sqrt{10}$  and  $\sigma_2 = \sqrt{10}$ , with  $\sigma_1 > \sigma_2$ .

**3.E.7.** We see that  $A^T A$  is a diagonal matrix:

$$A^T A = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}.$$

From this we deduce that  $A^T A$  has three eigenvalues,  $\lambda_1 = 1$ ,  $\lambda_2 = 1/4$  and  $\lambda_3 = 0$ . and two non-zero singular values (note that  $\text{rank}(A) = 2$ ), the latter given by  $\sigma_1 = \sqrt{\lambda_1} = 1$  and  $\sigma_2 = \sqrt{\lambda_2} = 1/2$ , with  $\sigma_1 > \sigma_2$ . Therefore, the matrix  $\Sigma$  has the form  $\Sigma = \text{diag}(1, 1/2, 0)$ .

**3.E.9.** According to 3.5.5 we will have  $P := U\Sigma U^T$  and  $U_p := UV^T$ , where  $A = U\Sigma V^T$  is the SVD decomposition of  $A$ . We compute

$$P = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$U_p = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}.$$

Thus, the polar decomposition of the matrix  $A$  has the form

$$A = \begin{pmatrix} 0 & 0 & -\frac{1}{2} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = PU_p = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}.$$

A verification using Sage is easy:

```
A=matrix([[0, 0, -1/2], [-1, 0, 0], [0, 0, 0]])
S=diagonal_matrix([1, 1/2, 0])
V=matrix([[1, 0, 0], [0, 0, 1], [0, 1, 0]])
U=matrix([[0, -1, 0], [-1, 0, 0], [0, 0, -1]])
P=U*S*(U.T)
show(P)
Upolar=U*(V.T)
show(Upolar)
A==P*Upolar
```

The polar decomposition is not unique, since  $A$  is singular. Next it easy to see that  $P$  is Hermitian and positive semi-definite. For example, for the second claim we see that  $P$  has three non-negative eigenvalues: 1, 1/2 and 0. In Sage you can confirm these claims by the adding the cell

```
show(P.eigenvalues())
P.is_hermitian()
```

**3.F.10.** The answer is given by  $x_n = (-1)^n(-2n^2 + 8n - 7)$ .

**3.F.23.** The required matrix has the form  $\begin{pmatrix} \frac{5}{6} & \frac{1}{5} \\ \frac{1}{6} & \frac{5}{5} \end{pmatrix}$ . This matrix has the dominant eigenvalue 1, with corresponding eigenvector given by  $(\frac{6}{5}, 1)^T$ . Because the eigenvalue is dominant, the ratio of the viewers stabilises on 6 : 5.

**3.F.25.** The game ends after three bets (the same as in 3.F.24). Thus, all the powers of  $A$ , starting with  $A^3$ , are identical. Consequently the answer is given as follows:

$$A^{100} = A^3 = \begin{pmatrix} 1 & 7/8 & 3/4 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1/8 & 1/4 & 1/2 & 1 \end{pmatrix}.$$

**3.F.31.** It is easy to see that  $F$  is linear. Also we compute  $\ker(F) = \left\{ \begin{pmatrix} a & -a \\ a & -a \end{pmatrix} : a \in \mathbb{R} \right\}$  hence  $\dim_{\mathbb{R}}(\ker(F)) = 1$  and a basis of  $\ker(F)$  consists of the matrix  $u = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$ . Since  $\mathcal{B}(u, u) = 4$ , we deduce that  $\{\hat{u} = \frac{1}{2}u\}$  is an orthonormal basis of  $\ker(F)$ .

**3.F.36.** The given rotation is easily obtained by composing the following three mappings:

- rotation through the angle  $\pi/4$  in the negative sense about the axis  $z$  (the axis of the rotation goes over on the  $x$  axis);
- rotation through the angle  $\pi/3$  in the positive sense about the  $x$  axis;
- rotation through the angle  $\pi/4$  in the positive sense about the  $z$  axis (the  $x$  axis goes over on the axis of the rotation).

The matrix of the resulting rotation is the product of the matrices corresponding to the given three mappings, while the order of the matrices is given by the order of application of the mappings – the first mapping applied is in the product the rightmost one. Thus we obtain the desired matrix

$$\begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & \frac{\sqrt{6}}{4} \\ \frac{1}{4} & \frac{3}{4} & -\frac{\sqrt{6}}{4} \\ -\frac{\sqrt{6}}{4} & \frac{\sqrt{6}}{4} & \frac{1}{2} \end{pmatrix}.$$

Note that the resulting rotation could be also obtained for instance by taking the composition of the three following mappings:

- rotation through the angle  $\pi/4$  in the positive sense about the axis  $z$  (the axis of rotation goes over on the axis  $y$ );
- rotation through the angle  $\pi/3$  in the positive sense about the axis  $y$ ;
- rotation through the angle  $\pi/4$  in the negative sense about the axis  $z$  (the axis  $y$  goes over to the axis of rotation).

Analogously we obtain

$$\begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ -\frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & \frac{\sqrt{6}}{4} \\ \frac{1}{4} & \frac{3}{4} & -\frac{\sqrt{6}}{4} \\ -\frac{\sqrt{6}}{4} & \frac{\sqrt{6}}{4} & \frac{1}{2} \end{pmatrix}.$$

**3.F.39.** For any  $x, y \in \mathbb{R}^n$  and with respect to the standard dot product on  $\mathbb{R}^n$  we compute

$$\langle L_A x, y \rangle = \langle Ax, y \rangle = y^T Ax = y^T (A^T)^T x = (A^T y)^T x = \langle x, A^T y \rangle = \langle x, Ay \rangle$$

where the last equality follows since by assumption  $A = A^T$ .

**3.F.40.** First notice that the operator  $L_A : \text{Mat}_{m,n}(\mathbb{F}) \rightarrow \text{Mat}_{m,n}(\mathbb{F})$  is linear since  $L_A(B+C) = A(B+C) = AB+AC = L_A(B) + L_A(C)$  and  $L_A(zB) = A(zB) = z(AB) = zL_A(B)$  for any  $B, C \in \text{Mat}_{m,n}(\mathbb{F})$  and  $z \in \mathbb{F}$ . Its adjoint operator  $L_A^*$  is the linear map  $L_A^* : \text{Mat}_{m,n}(\mathbb{F}) \rightarrow \text{Mat}_{m,n}(\mathbb{F})$  satisfying

$$\mathcal{B}(L_A(B), C) = \mathcal{B}(B, L_A^*(C)), \quad B, C \in \text{Mat}_{m,n}(\mathbb{F}).$$

We will show that for  $L_A^* = A^*$  this holds like an identity. We compute

$$\mathcal{B}(L_A(B), C) = \text{tr}(C^*(AB)) = \text{tr}(C^*AB),$$

$$\mathcal{B}(B, L_A^*(C)) = \mathcal{B}(B, A^*C) = \text{tr}((A^*C)^*B) = \text{tr}((C^*(A^*)^*)B) = \text{tr}(C^*AB).$$

Since  $L_A^* = A^*$  and  $A$  is by assumption Hermitian, our claim follows.

**3.F.41.** Let  $y \in V$ . Then we have the following equivalences:

$$\begin{aligned} y \in \ker(f^*) &\iff f^*(y) = 0, \\ &\iff \langle x, f^*(y) \rangle = 0, \text{ for any } x \in V, \\ &\iff \langle f(x), y \rangle = 0, \text{ for any } x \in V, \\ &\iff y \perp f(x), \text{ for any } x \in V, \\ &\iff y \in (\text{Im}(f))^\perp. \end{aligned}$$

Thus  $\ker(f^*) = (\text{Im}(f))^\perp$ . Similarly is treated the second relation.

**3.F.42.** Reflections matrices are unitary since they preserve the length of vectors. In particular,

$$[\varphi_u][\varphi_u]^* = (2uu^T - E)(2uu^T - E)^* = (2uu^T - E)(2uu^T - E) = 4u(u^T u)u^T - 4uu^T + E = 4uu^T - 4uu^T + E = E,$$

since  $u$  is a unit vector, i.e.,  $\langle u, u \rangle = u^T u = 1$ .

**3.F.43.** Observe that  $(A^*A)^* = A^*A$  hence the matrix  $B = A^*A$  is Hermitian,  $B = B^*$ . Then the result follows since any Hermitian matrix has real eigenvalues, see 3.4.7. For the second statement, we have  $AA^* = U^*DU(U^*DU)^* = U^*DUU^*D^*U = U^*DD^*U$  and  $A^*A = (U^*DU)^*U^*DU = U^*D^*UU^*DU = U^*D^*DU$ . Now the result follows since  $D$  is diagonal and hence we have  $DD^* = D^*D$  (as one can easily verify).

**3.F.44.** It should be  $a = 2 + 2i$ ,  $b = 1 - i$  and  $c = i$ . One can verify this answer in Sage as follows:

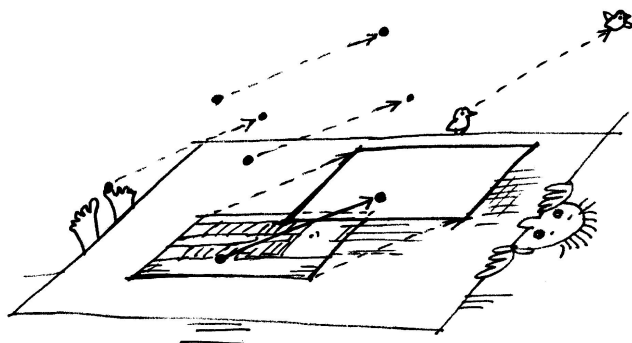
```
A=matrix(SR, 3, 3, [[1, 2+2*I, -1], [2-2*I, 0, 1-I], [1, 1+I, 0]])
A.is_hermitian()
```

Could you suggest an alternative method in Sage to verify the same result?

**3.F.51.** The answer is  $\dim \text{Mat}_n(\mathbb{R}) = n^2$ .

## Analytic geometry

*position, incidence, projection*  
 – and we return to matrices again...



### A. Affine geometry

**4.A.1.** Find a parametric equation for a line in  $\mathbb{R}^3$  given by the equations



$$\begin{aligned} x - 2y + z &= 2, \\ 2x + y - z &= 5. \end{aligned}$$

**Solution.** It is sufficient to solve the equation system. However, there is an alternative approach. Find a non-zero direction vector orthogonal to the normal vectors  $(1, -2, 1)$ ,  $(2, 1, -1)$ . The cross product

$$(1, -2, 1) \times (2, 1, -1) = (1, 3, 5)$$

is such a vector. The triple

$$[x, y, z] = [2, -1, -2]$$

satisfies the respective system, so a solution is

$$[2, -1, -2] + t(1, 3, 5), \quad t \in \mathbb{R}.$$

□

**4.A.2.** A plane in  $\mathbb{R}^4$  is given by its parametric equation

$$\varrho : [0, 3, 2, 5] + t(1, 0, 1, 0) + s(2, -1, -2, 2), \quad t, s \in \mathbb{R}$$

Find its implicit equation.

**Solution.** The task is to find a system of equations with 4 variables  $x, y, z, u$  (the dimension of the space is 4) which are satisfied by the coordinates of those points which lie in the plane. The desired system must contain  $2 = 4 - 2$  linearly

We return to the view on geometry that we had when we studied positions of points in the plane in the 5th part of the first chapter, c.f. 1.5.1. We are interested in the properties of objects in the Euclidean space, delimited by points, straight lines, planes etc. The essential point is to clarify how their properties are related to the notion of vectors, and whether they depend on the notion of the length of vectors.

Later in the book, we use linear and multi-linear algebra to study objects defined in a nonlinear way. In order to facilitate this we provide some insight in the so called analytic geometry, based on the application of matrix calculus. This point of view is most useful in discussion of the technique of optimization, e.g. when searching for (constrained) extrema of functions.

At the end of this chapter we show how the projectivization of affine spaces helps us to obtain simplification and stability of algorithms typical for computer graphics.

### 1. Affine and Euclidean geometry

While clarifying the structure of solutions of linear equations in the second chapter we find in paragraph 2.3.5 that the set of all solutions of a nonhomogeneous system of linear equations does not form a vector space. However, the solutions always arise in such a way that to one particular solution we can add the vector space of solutions to the corresponding homogeneous system. On the other hand, the difference of any two solutions of the nonhomogeneous system is always a solution of the homogeneous system. This behaviour is similar to the behaviour of linear difference equations. We see this already in paragraph 3.2.6.



**4.1.1. Affine spaces.** Going back to paragraph 1.5.3 (and further on) provides a hint how to deal with the theory in any finite dimension. There we describe straight lines and points as sets of solutions of systems of linear equations. A line is considered as a one-dimensional subspace, although its points are described by two coordinates. Parametrically, the line is defined by the sum of a single point (that is, a pair of coordinates) and multiples of a fixed direction vector. We proceed now in the same way for arbitrary dimensions.

independent equations. Solve the problem by elimination of parameters. The points  $[x, y, z, u] \in \varrho$  satisfy

$$\begin{aligned} x &= t + 2s, \\ y &= 3 - s, \\ z &= 2 + t - 2s, \\ u &= 5 + 2s, \end{aligned}$$

where  $t, s \in \mathbb{R}$ . Write the system as a matrix

$$\left( \begin{array}{cc|cccc|c} 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 3 \\ 1 & -2 & 0 & 0 & -1 & 0 & 2 \\ 0 & 2 & 0 & 0 & 0 & -1 & 5 \end{array} \right).$$

The first two columns are direction vectors of the plane, followed by the negative identity matrix. The last column is the vector of coordinates of the point  $[0, 3, 2, 5]$ . This is now a system in  $t, s, x, y, z, u$ . Transform the obtained matrix using elementary row operations in order to have as many zero-rows on the left-hand side of the first vertical line as possible. Adding  $(-1)$ -times the first row and  $(-4)$ -times the second row to the third row and adding twice the second row to the first row gives

$$\left( \begin{array}{cc|cccc|c} 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 & 3 \\ 0 & 0 & 1 & 4 & -1 & 0 & -10 \\ 0 & 0 & 0 & -2 & 0 & -1 & 11 \end{array} \right).$$

The bottom two rows, both with only zeros to the left of the first vertical line, imply

$$\begin{aligned} x + 4y - z - 10 &= 0, \\ -2y - u + 11 &= 0. \end{aligned}$$

Note that the original system can be written as

$$\left( \begin{array}{cccc|cc|c} 1 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 3 \\ 0 & 0 & 1 & 0 & 1 & -2 & 2 \\ 0 & 0 & 0 & 1 & 0 & 2 & 5 \end{array} \right),$$

where  $x, y, z, u$  remains on the left-hand side of the equations. A similar transformation gives

$$\left( \begin{array}{cccc|cc|c} 1 & 0 & 0 & 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 3 \\ -1 & -4 & 1 & 0 & 0 & 0 & -10 \\ 0 & 2 & 0 & 1 & 0 & 0 & 11 \end{array} \right)$$

from which

$$\begin{aligned} -x - 4y + z &= -10, \\ 2y + u &= 11. \end{aligned}$$

As seen in this exercise, parameter elimination can be long-winded. It is not difficult to make a mistake along the way.

**Another solution** All that is needed are two linearly independent vectors perpendicular to  $(1, 0, 1, 0)$ ,  $(2, -1, -2, 2)$ . If we “guessed” that these vectors could be for example  $(0, 2, 0, 1)$ ,  $(-1, 0, 1, 2)$ , then putting  $x = 0$ ,  $y = 3$ ,  $z = 2$ ,  $u = 5$  to the equations

$$\begin{aligned} 2y + u &= a, \\ -x + z + 2u &= b \end{aligned}$$

STANDARD AFFINE SPACE

*Standard affine space*  $\mathcal{A}_n$  is a set of all points in  $\mathbb{R}^n = \mathcal{A}_n$  together with an operation “+” which assigns the point

$$A + v = (a_1 + v_1, \dots, a_n + v_n) \in \mathbb{R}^n = \mathcal{A}_n$$

to a point  $A = (a_1, \dots, a_n) \in \mathcal{A}_n$  and a vector  $v = (v_1, \dots, v_n) \in \mathbb{R}^n = V$ .

This operation satisfies the following three properties:

- (1)  $A + 0 = A$  for all points  $A \in \mathcal{A}_n$  and the null vector  $0 \in V$ ,
- (2)  $A + (v + w) = (A + v) + w$  for all vectors  $v, w \in V$  and points  $A \in \mathcal{A}_n$ ,
- (3) for every two points  $A, B \in \mathcal{A}_n$  there exists exactly one vector  $v \in V$  such that  $A + v = B$ . This vector is denoted by  $v = B - A$ , sometimes also  $\overline{AB}$ .

The underlying vector space  $\mathbb{R}^n$  is called the *difference space* of the standard affine space  $\mathcal{A}_n$ .



Notice that care is needed about several formal ambiguities. In particular, the symbol “+” is used for two different operations. “+” is used for adding a vector from the difference space to a point in the affine space. “+” is also used for summing vectors in the difference space  $V = \mathbb{R}^n$ . Further, notice that the operation “-” assigns a vector to a couple of points and cannot be understood as inverse of the addition (although in the standard affine space it is given by the difference of the  $n$ -tuples  $B$  and  $A$ ). We do not introduce specific letters for the set of points in the affine space.  $\mathcal{A}_n$  denotes both this set of points as well as the whole structure defining the affine space.

Why distinguish between the set of points in the affine space  $\mathcal{A}_n$  and its difference space  $V$  when both spaces can be viewed as  $\mathbb{R}^n$ ? It is a fundamental formal step to understanding the geometry in  $\mathbb{R}^n$ : The issue is that geometric objects, namely straight lines, points, planes etc. do not depend directly on the vector space structure of the set  $\mathbb{R}^n$ . They do not depend at all on the fact that we work with  $n$ -tuples of scalars. We need to know only what it means to move “straight in a given direction”. For instance, we can consider the affine plane as an unbounded board without chosen coordinates, but with the possibility of moving about a given vector. When we switch to such an abstract view, we can discuss the “plane geometry” for two-dimensional subspaces, without the need to work with  $k$ -tuples of coordinates.

This point of view underlies the following definition:

**4.1.2. Definition.** The affine space  $\mathcal{A}$  with the difference space  $V$  is a set of *points*  $\mathcal{P}$ , together with the map

$$\mathcal{P} \times V \rightarrow \mathcal{P}, \quad (A, v) \mapsto A + v.$$

$V$  is a vector space. The map satisfies the properties (1)–(3) from the definition of the standard affine space.

For a fixed vector  $v \in V$ , there is a *translation*  $\tau_v : \mathcal{A} \rightarrow \mathcal{A}$  as the restricted map

$$\tau_v : \mathcal{P} \simeq \mathcal{P} \times \{v\} \rightarrow \mathcal{P}, \quad A \mapsto A + v.$$

yields  $a = 11, b = 12$ . The desired implicit expression is

$$\begin{aligned} 2y &+ u = 11, \\ -x &+ z + 2u = 12. \end{aligned}$$

**Another solution** Since

$$\begin{aligned} x &= t + 2s, \\ y &= 3 - s, \\ z &= 2 + t - 2s, \\ u &= 5 + 2s, \end{aligned}$$

Eliminate  $t$  to get

$$\begin{aligned} x - z &= 2 - 4s, \\ y &= 3 - s, \\ u &= 5 + 2s, \end{aligned}$$

Eliminate  $s$  to obtain two equations, namely

$$\begin{aligned} z - x + 2u &= 12 \\ u + 2y &= 11, \end{aligned}$$

which solves the problem.  $\square$

**4.A.3.** Find a parametric equation of the plane passing through the points

$$A = [2, 1, 1], \quad B = [3, 4, 5], \quad C = [4, -2, 3].$$

Hence find a parametric equation of the open half-plane containing the point  $C$  and bounded by the line passing through the points  $A, B$ .

**Solution.** We need one point and two (linearly independent) vectors lying in the plane. It is enough to choose  $A$  together with the vectors  $B - A = (1, 3, 4)$  and  $C - A = (2, -3, 2)$ , which are clearly independent. A point  $[x, y, z]$  lies in the plane if and only if there exist numbers  $t, s \in \mathbb{R}$  so that

$$x = 2 + 1 \cdot t + 2 \cdot s, \quad y = 1 + 3 \cdot t - 3 \cdot s, \quad z = 1 + 4 \cdot t + 2 \cdot s.$$

Consequently a parametric equation is

$$[x, y, z] = [2, 1, 1] + t(1, 3, 4) + s(2, -3, 2), \quad t, s \in \mathbb{R}.$$

Setting  $s = 0$  gives a line passing through the points  $A$  and  $B$ .  $t = 0$  and  $s \geq 0$ , defines a ray passing through  $C$  with an initial point  $A$ . A particular but arbitrarily chosen  $t \in \mathbb{R}$  and variable  $s \geq 0$  gives a ray initiated on the border line, going through the half-plane in which the point  $C$  lies. That means that the desired open half-plane can be expressed parametrically as

$$[x, y, z] = [2, 1, 1] + t(1, 3, 4) + s(2, -3, 2), \quad t \in \mathbb{R}, s > 0. \quad \square$$

**4.A.4.** Determine the relative position of the lines

$$\begin{aligned} p &: [1, 0, 3] + t(2, -1, -3), \quad t \in \mathbb{R}, \\ q &: [1, 1, 3] + s(1, -1, -2), \quad s \in \mathbb{R}. \end{aligned}$$

**Solution.** Search for common points of the given lines (subspaces intersection). We have a system

$$\begin{aligned} 1 + 2t &= 1 + s, \\ 0 - t &= 1 - s, \\ 3 - 3t &= 3 - 2s. \end{aligned}$$

By the *dimension* of an affine space  $\mathcal{A}$ , is meant the dimension of its difference space.

In the sequel, we do not distinguish accurately between  $\mathcal{A}$  and the set of points  $\mathcal{P}$ . We talk instead about points and vectors of the affine space  $\mathcal{A}$ .

It follows immediately from the axioms that for arbitrary points  $A, B, C$  in the affine space  $\mathcal{A}$

$$\begin{aligned} (1) \quad A - A &= 0 \in V \\ (2) \quad B - A &= -(A - B) \\ (3) \quad (C - B) + (B - A) &= C - A. \end{aligned}$$

Indeed, (1) follows from the fact that  $A + 0 = A$  and that such a vector is unique (the first and third defining property). By adding successively  $B - A$  and  $A - B$  to  $A$ , according to the second defining property we obtain  $A$  again. Add the null vector to prove (2). Similarly, (3) follows from the defining property 4.1.1 (2) and the uniqueness.

Notice that the choice of one fixed point  $A_0 \in \mathcal{A}$  determines a bijection between  $V$  and  $\mathcal{A}$ . So for a fixed basis  $\underline{u}$  in  $V$  there is a unique expression

$$A = A_0 + x_1 u_1 + \cdots + x_n u_n$$

for every point  $A \in \mathcal{A}$ . We talk about an *affine coordinate system*  $(A_0; u_1, \dots, u_n)$  given by the *origin of the affine coordinate system*  $A_0$  and the basis  $\underline{u}$  of the corresponding difference space. This is sometimes called an *affine frame*  $(A_0, \underline{u})$ .

To summarize: Affine coordinates of a point  $A$  in the frame  $(A_0, \underline{u})$  are the coordinates of the vector  $A - A_0$  in the basis  $\underline{u}$  of the difference space  $V$ .

The choice of an affine coordinate system identifies each  $n$ -dimensional affine space  $\mathcal{A}$  with the standard affine space  $\mathcal{A}_n$ .

**4.1.3. Affine subspaces.** If we consider some coordinate system on  $\mathcal{A}$  and choose only such points in  $\mathcal{A}$  which have some chosen coordinates equal to zero (for instance the last one), we obtain again a set which behaves as an affine space. This is the spirit of the following definition of the affine subspaces.



SUBSPACES OF AN AFFINE SPACE

**Definition.** The nonempty subset  $\mathcal{Q} \subset \mathcal{A}$  of an affine space  $\mathcal{A}$  with a difference space  $V$  is called an *affine subspace* in  $\mathcal{A}$  if the subset  $W = \{B - A; A, B \in \mathcal{Q}\} \subset V$  is a vector subspace and  $A + v \in \mathcal{Q}$  for any  $A \in \mathcal{Q}, v \in W$ .

It is important to include both of the conditions in the definition, since there are examples of sets which satisfy the first condition but not the second. One such set consists of a straight line in the plane with one point removed.

For an arbitrary set of points  $M \subset \mathcal{A}$  in an affine space with a difference space  $V$ , we define the vector space

$$Z(M) = \text{span}\{B - A; B, A \in M\} \subset V$$

of all vectors generated by the differences of points in  $M$ .

From the first two equations,  $t = 1, s = 2$ . This does not satisfy the third equation. Thus the system does not have a solution. The direction vector  $(2, -1, -3)$  of the line  $p$  is not a multiple of the direction vector  $(1, -1, -2)$  of the line  $q$ . Hence the lines are not parallel. Hence, the lines are skew.  $\square$

**4.A.5.** Find all numbers  $a \in \mathbb{R}$  so that the lines

$$p : [4, -4, 8] + t(2, 1, -4), \quad t \in \mathbb{R},$$

$$q : [a, 6, -5] + s(1, -3, 3), \quad s \in \mathbb{R}$$

intersect.

**Solution.** The lines intersect if and only if the system

$$\begin{aligned} 4 + 2t &= a + s, \\ -4 + t &= 6 - 3s, \\ 8 - 4t &= -5 + 3s \end{aligned}$$

has a solution. Express the system as a matrix (the first column corresponding to  $t$ , the second to  $s$ ), and solve

$$\begin{aligned} \left( \begin{array}{cc|c} 2 & -1 & a-4 \\ 1 & 3 & 10 \\ -4 & -3 & -13 \end{array} \right) &\sim \left( \begin{array}{cc|c} 1 & 3 & 10 \\ 2 & -1 & a-4 \\ -4 & -3 & -13 \end{array} \right) \\ &\sim \left( \begin{array}{cc|c} 1 & 3 & 10 \\ 0 & -7 & a-24 \\ 0 & 1 & 3 \end{array} \right). \end{aligned}$$

The system has a solution if and only if the second row is a multiple of the third row. This property is satisfied only for  $a = 3$ . The point of intersection of the lines is  $[6, -3, 4]$ .  $\square$

**4.A.6.** In  $\mathbb{R}^3$ , determine the relative position of the line  $p$  defined implicitly by

$$\begin{aligned} x + y - z &= 4, \\ x - 2y + z &= -3 \end{aligned}$$

and the plane  $\varrho : y = 2x - 1$ .

**Solution.** A normal vector to the plane is  $\varrho$  is  $(2, -1, 0)$  (consider  $\varrho : 2x - y + 0z = 1$ ). Since

$$(1, 1, -1) + (1, -2, 1) = (2, -1, 0),$$

the normal vector to the plane  $\varrho$  is a linear combination of the  $p$  normal vectors. A vector defining the line lies in a subspace of the plane  $\varrho$ . It remains to discover whether or not they intersect. The system of equations

$$\begin{aligned} x + y - z &= 4, \\ x - 2y + z &= -3, \\ 2x - y &= 1 \end{aligned}$$

has infinitely many solutions, because the first two equations add to give the third one. So the line  $p$  lies in the plane  $\varrho$ .  $\square$

The following exercise is a typical vector spaces intersection exercise. The reader should be able to solve this. Otherwise we recommend not continuing with this book.



In particular,  $V = Z(\mathcal{A})$ . Every affine subspace  $\mathcal{Q} \subset \mathcal{A}$  itself satisfies the axioms for an affine space with the difference space  $Z(\mathcal{Q})$ .

The intersection of any set of affine subspaces is either an affine subspace or is the empty set. This follows directly from the definitions.

The affine subspace  $\langle M \rangle$  in  $\mathcal{A}$  generated by a nonempty set  $M \subset \mathcal{A}$  is the intersection of all affine subspaces which contain all points of  $M$ .

#### AFFINE HULL AND PARAMETRIC DESCRIPTION

Affine subspaces can be described by their difference spaces after choosing a point  $A_0 \in M$  in a generating set  $M$ . Indeed, to generate the affine subspace, take the vector subspace  $Z(M) \subset Z(\mathcal{A})$  generated by all differences of points in  $M$ , and add this vector space to an arbitrary point in  $M$ , i.e.,  $\langle M \rangle = \{A_0 + v; v \in Z(M) \subset Z(\mathcal{A})\}$ . We talk about the *affine hull* of the set of points  $M$  in  $\mathcal{A}$ .

On the other hand, whenever a subspace  $U$  in the difference space  $Z(\mathcal{A})$  and a fixed point  $A \in \mathcal{A}$  is chosen, the subset  $A + U$ , created by all possible sums of  $A$  and all vectors in  $U$ , is an affine subspace. This approach leads to the notion of parametrization of subspaces:

Let  $\mathcal{Q} = A + Z(\mathcal{Q})$  be an affine subspace in  $\mathcal{A}_n$ . Let  $(u_1, \dots, u_k)$  be a basis of  $Z(\mathcal{Q}) \subset \mathbb{R}^n$ . Then the expression of the subspace

$$\mathcal{Q} = \{A + t_1u_1 + \dots + t_ku_k; t_1, \dots, t_k \in \mathbb{R}\}$$

is called the *parametric description* of the subspace  $\mathcal{Q}$ .

In the sequel, we shall use the same notation  $\langle \rangle$  for the affine hulls in the affine spaces and linear hulls in their difference vector spaces.

There is another way of prescribing affine spaces: If we choose affine coordinates, then the difference space may be described by a homogeneous system of linear equations in these coordinates. By inserting the coordinates of one point of the subspace  $\mathcal{Q}$  into the system of equations, we obtain the right-hand side of the non-homogeneous system with the same matrix. The subspace  $\mathcal{Q}$  is exactly the set of solutions of this system. The description of the subspace  $\mathcal{Q}$  by a system of equations in given coordinates is called an *implicit description* of the subspace  $\mathcal{Q}$ .

The following proposition says that we can prescribe all affine subspaces in this way. It shows the geometric nature of the solutions of systems of linear equations.

**4.1.4. Theorem.** Let  $(A_0; \underline{u})$  be an affine coordinate system in an  $n$ -dimensional affine space  $\mathcal{A}$ . In these coordinates, affine subspaces of dimension  $k$  in  $\mathcal{A}$  are exactly the sets of solutions of solvable systems of  $n - k$  linearly independent equations in  $n$  variables.

**PROOF.** Consider an arbitrary solvable system of  $n - k$  linearly independent equations  $\alpha_i(x) = b_i$ , where  $b_i \in \mathbb{R}$ ,  $i = 1, \dots, n - k$ . Suppose  $A = (a_1, \dots, a_n)^T \in \mathbb{R}^n$  is a

**4.A.7.** Find the intersection of the subspaces  $Q_1$  and  $Q_2$ , where

$$Q_1 : [4, -5, 1, -2] + t_1(3, 5, 4, 2) + t_2(2, 4, 5, 1) + t_3(0, 3, 1, 2),$$

$$Q_2 : [4, 4, 4, 4] + s_1(0, -6, -2, -4) + s_2(-1, -5, -3, -3),$$

for  $t_1, t_2, t_3, s_1, s_2 \in \mathbb{R}$ .

**Solution.** The point  $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^4$  lies in  $Q_1$  if and only if

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ -5 \\ 1 \\ -2 \end{bmatrix} + t_1 \begin{bmatrix} 3 \\ 5 \\ 4 \\ 2 \end{bmatrix} + t_2 \begin{bmatrix} 2 \\ 4 \\ 5 \\ 1 \end{bmatrix} + t_3 \begin{bmatrix} 0 \\ 3 \\ 1 \\ 2 \end{bmatrix}$$

for some numbers  $t_1, t_2, t_3 \in \mathbb{R}$ . The point  $X = [x_1, x_2, x_3, x_4] \in \mathbb{R}^4$  lies in  $Q_2$  if and only if

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + s_1 \begin{bmatrix} 0 \\ -6 \\ -2 \\ -4 \end{bmatrix} + s_2 \begin{bmatrix} -1 \\ -5 \\ -3 \\ -3 \end{bmatrix}$$

for some  $s_1, s_2 \in \mathbb{R}$ . Hence if  $X$  lies in  $Q_1 \cap Q_2$  then the equation

$$t_1 \begin{bmatrix} 3 \\ 5 \\ 4 \\ 2 \end{bmatrix} + t_2 \begin{bmatrix} 2 \\ 4 \\ 5 \\ 1 \end{bmatrix} + t_3 \begin{bmatrix} 0 \\ 3 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 4-4 \\ 4+5 \\ 4-1 \\ 4+2 \end{bmatrix} + s_1 \begin{bmatrix} 0 \\ -6 \\ -2 \\ -4 \end{bmatrix} + s_2 \begin{bmatrix} -1 \\ -5 \\ -3 \\ -3 \end{bmatrix}.$$

has a solution for  $t_1, t_2, t_3, s_1, s_2$ .

Move the vectors corresponding to  $s_1$  and  $s_2$  to the left-hand side. Write the equations in matrix form and reduce to echelon form. There follows

$$\begin{pmatrix} 3 & 2 & 0 & 0 & 1 & 0 \\ 5 & 4 & 3 & 6 & 5 & 9 \\ 4 & 5 & 1 & 2 & 3 & 3 \\ 2 & 1 & 2 & 4 & 3 & 6 \end{pmatrix} \sim \begin{pmatrix} 3 & 2 & 0 & 0 & 1 & 0 \\ 0 & 2 & 9 & 18 & 10 & 27 \\ 0 & 7 & 3 & 6 & 5 & 9 \\ 0 & -1 & 6 & 12 & 7 & 18 \end{pmatrix}$$

$$\sim \dots \sim \left( \begin{array}{cccccc|c} 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array} \right).$$

So  $t_1 = t_2 = s_2 = 0$  and for  $s_1 = t \in \mathbb{R}$  we have  $t_3 = 3 - 2t$ . Note that for the determination of  $Q_1 \cap Q_2$ , it is sufficient to know either  $t_1, t_2, t_3$ , or  $s_1, s_2$ . So

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + s_1 \begin{bmatrix} 0 \\ -6 \\ -2 \\ -4 \end{bmatrix} + s_2 \begin{bmatrix} -1 \\ -5 \\ -3 \\ -3 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix} + t \begin{bmatrix} 0 \\ -6 \\ -2 \\ -4 \end{bmatrix}.$$

This should be checked using  $t_1 = t_2 = 0$  and  $t_3 = 3 - 2t$ . The solution is a line lying in both planes. It is  $Q_1 \cap Q_2$ .  $\square$

fixed solution of this (non-homogeneous) system. Suppose also that  $U \subset \mathbb{R}^n$  is the vector space of all solutions of the homogenized system  $\alpha_i(x) = 0$ . Then the dimension of  $U$  is  $k$ . The set of all solutions of the given system is of the form  $\{B; B = A + (y_1, \dots, y_n)^T, y = (y_1, \dots, y_n)^T \in U\} \subset \mathbb{R}^n$ , c.f. 2.3.5. So the corresponding affine subspace is described parametrically by the initial coordinates  $(A_0; \underline{u})$ .

Conversely, consider an arbitrary affine subspace  $Q \subset \mathbb{A}^n$ . Choose a point  $B$  therein, and consider this point to be the origin of an affine coordinate system  $(B, \underline{v})$  for the affine space  $\mathcal{A}$ . Since  $Q = B + Z(Q)$ , it is necessary to describe the difference space of the subspace  $Q$  as a subspace of solutions of a homogeneous system of linear equations. Therefore, choose a basis  $\underline{v}$  of  $Z(\mathcal{A})$  such that the first  $k$  vectors form a basis of  $Z(Q)$ . In these coordinates, the vectors  $v \in Z(Q)$  are given by equations

$$\alpha_j(v) = 0, \quad j = k + 1, \dots, n.$$

The  $\alpha_i$  are linear forms from the dual basis to  $\underline{v}$ . They are the functions which assign to a vector the corresponding coordinates in the basis  $\underline{v}$ .

Hence the vector subspace  $Z(Q)$  of dimension  $k$  in the  $n$ -dimensional space  $\mathbb{R}^n$  is given as a solution of a homogeneous system of  $n - k$  independent equations. The description of the chosen affine subspace in the newly chosen coordinate system  $(B; \underline{v})$  is therefore given by a system of homogeneous linear equations.

It remains to consider the consequences of the transition from the former coordinate system  $(A; \underline{u})$  to the new adapted system  $(B; \underline{v})$ . It follows from a general consideration about transformations of coordinates in the following paragraph that the final description of the subspace is again a system of linear equations. This time it is non-homogeneous in general.  $\square$

**4.1.5. Coordinate transformations.** Any two arbitrarily chosen affine coordinate systems  $(A_0, \underline{u})$ ,



$(B_0, \underline{v})$  differ in the basis of the difference spaces and the origin of the latter one is translated about the vector  $(B_0 - A_0)$ . Hence the equations for the corresponding coordinate transformations can be read off from the rule for a transformation of a point  $X \in \mathcal{A}$

$$X = B_0 + x'_1 v_1 + \dots + x'_n v_n$$

$$= B_0 + (A_0 - B_0) + x_1 u_1 + \dots + x_n u_n.$$

Let  $y = (y_1, \dots, y_n)^T$  denote the column of coordinates of the vector  $(A_0 - B_0)$  in the basis  $\underline{v}$ . Let  $M = (a_{ij})$  be the matrix expressing the basis  $\underline{u}$  in terms of the basis  $\underline{v}$ . Then

$$x'_1 = y_1 + a_{11}x_1 + \dots + a_{1n}x_n$$

$$\vdots$$

$$x'_n = y_n + a_{n1}x_1 + \dots + a_{nn}x_n.$$

In matrix notation

$$x' = y + M \cdot x.$$



**4.A.8.** Determine whether or not the points  $[0, 2, 1]$ ,  $[-1, 2, 0]$ ,  $[-2, 5, 2]$  and  $[0, 5, 4]$  in  $\mathbb{R}^3$  all lie in the same plane.

**Solution.** Consider the vectors  $[0, 2, 1] - [-1, 2, 0] = (1, 0, 1)$ ,  $[0, 2, 1] - [-2, 5, 2] = (2, -3, -1)$  and  $[0, 2, 1] - [0, 5, 4] = (0, -3, -3)$ . They are linearly dependent since the matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ 2 & -3 & -1 \\ 0 & -3 & -3 \end{pmatrix},$$

has rank 2. Hence, the given points lie in a plane.  $\square$

**4.A.9.** Into how many parts can three planes slice the space ( $\mathbb{R}^3$ )? Give an example of planes in a suitable position for every case.

**4.A.10.** Determine whether or not the point  $[2, 1, 0]$  lies within the convex hull of the points  $[0, 2, 1]$ ,  $[1, 0, 1]$ ,  $[3, -2, -1]$ ,  $[-1, 0, 1]$ .

**Solution.**  $[2, 1, 0]$  lies in the convex hull (see chapter 4.9) if and only if

$[2, 1, 0] = t_1[0, 2, 1] + t_2[1, 0, 1] + t_3[3, -2, -1] + t_4[-1, 0, 1]$  has a solution with  $t_1, t_2, t_3, t_4$ , all non-negative and  $t_1 + t_2 + t_3 + t_4 = 1$ . Equivalently,  $[2, 1, 0]$  lies in the convex hull if and only if

$$\begin{aligned} [2, 1, 0, 1] &= t_1[0, 2, 1, 1] + t_2[1, 0, 1, 1] + t_3[3, -2, -1, 1] \\ &\quad + t_4[-1, 0, 1, 1] \end{aligned}$$

has a solution with  $t_1, t_2, t_3, t_4$ , all non-negative. Solving these four equations, gives  $(t_1, t_2, t_3, t_4) = (1, 0, 1/2, -1/2)$ , so the given point does not lie in the convex hull.  $\square$

**4.A.11.** In  $\mathbb{R}^3$ , a tetrahedron has vertices  $ABCD$ , where  $A = [4, 0, 2]$ ,  $B = [-2, -3, 1]$ ,  $C = [1, -1, -3]$ ,  $D = [2, 4, -2]$ .

- Determine its volume.
- Decide whether or not the point  $X = [0, -3, 0]$  lies inside the tetrahedron.

**Solution.** a) The volume of the tetrahedron is one sixth of volume of a parallelepiped, of which three edges from the point  $A$  are  $B - A = (-6, -3, -1)$ ,  $C - A = (-3, -1, -5)$  and  $D - A = (-2, 4, -4)$ . It is given by the absolute value of the determinant

$$\begin{vmatrix} -6 & -3 & -1 \\ -3 & -1 & -5 \\ -2 & 4 & -4 \end{vmatrix} = -124.$$

Thus, the volume of the tetrahedron is  $\frac{124}{6}$ .

b) Write  $X$  as an affine combination of its vertices, by solving the system of four linear equation in four unknowns  $a, b, c, d$  given by the equality  $X = aA + bB + cC + dD$ . The solution is  $X = \frac{1}{4}A + \frac{1}{2}B + \frac{1}{2}C - \frac{1}{4}D$ . Since the coefficient of  $D$  is negative,  $X$  does not lie in the convex hull of the points  $A, B, C$  and  $D$ . Hence the given point does not lie inside the tetrahedron.  $\square$

Now, we can come back to the proof of the previous theorem. Suppose the system of linear equations in the “new” coordinates  $(B_0; \underline{v})$  has the form

$$S \cdot x' = b$$

where  $S$  is the matrix of the system. Then, in the original coordinates  $(A_0; \underline{u})$

$$S \cdot x' = S \cdot (y + M \cdot x) = b.$$

Thus in the original coordinates the system has the form

$$S \cdot M \cdot x = b - S \cdot y.$$

Therefore, if a subspace is described by a system of linear equations in one affine frame, then it is so described in all other affine frames. This completes the proof of the previous proposition.

**4.1.6. Examples of subspaces.** (1) The one-dimensional (standard) affine space is the set of all points of a real straight line  $\mathcal{A}_1$ . Its difference space is a one-dimensional vector space  $\mathbb{R}$ . The supporting set (i.e. the set of points in  $\mathcal{A}_1$ ) is also  $\mathbb{R}$ . The affine coordinates are obtained by the choice of an origin and a scale (i.e. a basis in the vector space  $\mathbb{R}$ ). All proper affine spaces are 0-dimensional. They are formed by all points of the real straight line  $\mathbb{R}$ .

(2) The two-dimensional (standard) affine space is a set of all points in the space  $\mathcal{A}_2$  with the difference space  $\mathbb{R}^2$ . The supporting set is  $\mathbb{R}^2$ . The affine coordinates are obtained by a choice of an origin and two linearly independent vectors (directions and scales). The proper subspaces are then all points and straight lines in the plane (0-dimensional and 1-dimensional). The lines are prescribed by the choice of a point and one vector from the corresponding difference space. The vector is a generator of direction as in the parametric definition of the straight line.

(3) The three-dimensional (standard) affine space is a set of all points in the space  $\mathcal{A}_3$  with the difference space  $\mathbb{R}^3$ . The affine coordinates are obtained by the choice of an origin and three linearly independent vectors (directions and scales). The proper affine subspaces are then all points, straight lines and planes (0-dimensional, 1-dimensional and 2-dimensional).

(4) Suppose there is given a nonzero vector of coefficients  $(a_1, \dots, a_n)$  and a scalar  $b \in \mathbb{R}$ . Compute the subspace of all solutions of one linear equation  $a \cdot x = b$  for the unknown point  $[x_1, \dots, x_n] \in \mathcal{A}_n$ . This is an affine subspace of dimension  $n - 1$ . We say that the subspace is of codimension 1, called a *hyperplane* in  $\mathcal{A}_n$ .

**4.1.7. Affine combinations of points.** We introduce an analogue of the linear combination of vectors. Let  $A_0, \dots, A_k$  be points in the affine space  $\mathcal{A}$ . Their affine hull  $\langle \{A_0, \dots, A_k\} \rangle$  can be written as the set of points

$$\{A_0 + t_1(A_1 - A_0) + \dots + t_k(A_k - A_0); t_1, \dots, t_k \in \mathbb{R}\}.$$

**4.A.12. Affine transformation of point coordinates**



The point  $X$  has coordinates expressed as  $[2, 2, 3]$  in an affine basis  $\{[1, 2, 3], (1, 1, 1), (1, -1, 2), (2, 1, 1)\}$  (in  $\mathbb{R}^3$ ). Determine its coordinates in the standard basis, i.e. in the basis  $\{[0, 0, 0], (1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ .

**Solution.** The coordinates  $[2, 2, 3]$  in the given basis are by the definition

$$\begin{aligned} & [1, 2, 3] + 2 \cdot (1, 1, 1) + 2 \cdot (1, -1, 2) + 3 \cdot (2, 1, 1) \\ &= [11, 5, 12] \end{aligned}$$

coordinates for  $X$  in the standard basis.  $\square$

**4.A.13. Affine transformation of mapping.** Find the affine mapping  $f$  in the coordinate system with the basis  $\underline{u} = \{(1, 1), (-1, 1)\}$  and origin  $[2, 0]$ , defined as

$$f(x_1, x_2) = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

in the standard basis in  $\mathbb{R}^2$ .

**Solution.** The change of the basis matrix from the basis  $\underline{u}$  to the standard basis is

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

The transformation matrix in the basis  $([2, 0], \underline{u})$  is obtained by first transforming the coordinates in the basis  $([2, 0], \underline{u})$  to the standard basis, i.e. to the basis  $([0, 0], (1, 0), (0, 1))$ . Then the transformation matrix  $f$  is applied in the standard basis. Finally it is transformed back to the coordinates in the basis  $([2, 0], \underline{u})$ . The transformation equations for changing the coordinates  $y_1, y_2$  in the basis  $([2, 0], \underline{u})$  to the coordinates  $x_1, x_2$  in the standard basis are

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

Hereby

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{-1} \left( \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right) \\ &= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}. \end{aligned}$$

Hence the desired mapping is

$$\begin{aligned} & f(y_1, y_2) \\ &= \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left[ \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \left( \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right) + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \end{pmatrix} \end{aligned}$$

In any affine coordinates the same set can be written as

$$\{t_0 A_0 + \dots + t_k A_k; t_i \in \mathbb{R}, \sum_{i=0}^k t_i = 1\}.$$

**AFFINE COMBINATIONS OF POINTS**

In general, by the formula  $t_0 A_0 + t_1 A_1 + \dots + t_k A_k$  with coefficients satisfying  $\sum_{i=0}^k t_i = 1$  is meant the set of all points  $A_0 + \sum_{i=1}^k t_i (A_i - A_0)$ . They are called the *affine combinations of points*.

The points  $A_0, \dots, A_k$  are in *general position* if they generate a  $k$ -dimensional affine subspace. This happens if and only if for each  $A_i$ , the vectors which arise as differences of this point  $A_i$ , and all other vectors  $A_j$ , are linearly independent. Observe that an assignment of a series of  $(\dim \mathcal{A}) + 1$  points in general position is equivalent to the definition of an affine frame with the origin in the first of them.

**4.1.8. Simplexes.** For points in an affine space, the affine combination is a similar construction to the linear combination for vectors in a vector space. Indeed, the affine subspace generated by points  $A_0, \dots, A_k$  equals the set of all affine combinations of its generators. The notion “to lie on the line between two points” can be generalized. In the two-dimensional case, imagine the interior of a triangle. In general proceed as follows:

**$k$ -DIMENSIONAL SIMPLEXES**

Let  $A_0, \dots, A_k$  be  $k + 1$  points in general position in an affine space  $\mathcal{A}$ . The set  $\Delta = \Delta(A_0, \dots, A_k)$  is defined as the set of all affine combinations of points  $A_i$  with nonnegative coefficients only. This is

$$\Delta = \{t_0 A_0 + t_1 A_1 + \dots + t_k A_k; 0 \leq t_i, \sum_{i=0}^k t_i = 1\},$$

called a  $k$ -dimensional *simplex* generated by the points  $A_i$ .

A one-dimensional simplex is a *line segment*, a two-dimensional simplex is a *triangle*, while a zero-dimensional simplex is a point.

Notice that each  $k$ -dimensional simplex has exactly  $k + 1$  *faces* defined by equations  $t_i = 0, i = 0, \dots, k$ . The faces are also simplexes, and their dimension is  $k - 1$ . We talk about the *boundary* of the simplex. For instance, the boundary of a triangle is formed by the three edges, and the boundary of each edge is formed by the two vertices.

The description of a subspace as a set of affine combinations of points in general position is equivalent to the parametric description. We work similarly with the parametric description of simplexes.

**4.1.9. Convex sets.** The subset  $M$  of an affine space is called *convex* if and only if for any two points  $A, B \in M$  the set contains the line segment  $\Delta(A, B)$ . Directly from the definition, each convex set with  $k + 1$  points in general position contains  $\square$  also the entire simplex defined by these points

**4.A.14.** Let there be a standard coordinate system in  $\mathbb{R}^3$  space. Agent K lives at the point  $S$  with coordinates  $[0, 1, 2]$ . The headquarters gave him a coordinate system with origin  $S$  and basis  $\{(1, 1, 0), (-1, 0, 1), (0, 1, 2)\}$ . Agent Bond lives at the point  $D$  with coordinates  $[1, 1, 1]$  and uses a coordinate system with basis  $\{(0, 0, 1), (-1, 1, 2), (1, 0, 1)\}$ . Agent K has set an appointment with agent Bond in the old brickfield which is (according to K's coordinate system) at the point  $[1, 1, 0]$ . To where should Bond go (regarding his coordinate system)?

**Solution.** The change of basis matrix from agent K's basis to the Bond's basis (with the same origins) is

$$T = \begin{pmatrix} -4 & 2 & -1 \\ 1 & 0 & 1 \\ 2 & -1 & 1 \end{pmatrix}$$

The vector  $(0, 1, 2)$  thus has coordinates  $T \cdot (0, 1, 2)^T = (0, 2, 1)^T$ . Translate the origin (add the vector  $(-1, 0, 1)$ ) to obtain the result  $(-1, 2, 2)$ .  $\square$

**4.A.15.** Find a transversal of the lines (that is, a line passing through both given lines)



$$[1, 1, 1] + t(2, 1, 0), \quad q: [2, 2, 0] + t(1, 1, 1),$$

so that  $[1, 0, 0]$  lies on the transversal.

**Solution.** The transversal lies in the plane  $\rho$  defined by the point  $[1, 0, 0]$  and the line  $p$ . Hence it lies in the plane

$$[1, 1, 1] + t(2, 1, 0) + s(0, 1, 1).$$

Let the point  $Q$  be the intersection of this plane with the line  $q$ .  $Q$  is obtained by solving the system

$$\begin{aligned} 1 + 2t &= 2 + u \\ 1 + t + s &= 2 + u \\ 1 + s &= u \end{aligned}$$

The left-hand sides of the equations represent all three coordinates of an arbitrary point of the plane  $\rho$  respectively. The right-hand sides then represent the coordinates of an arbitrary point on  $q$  (the free variable is denoted  $u$  in order not to be ambiguous). Solving this system, yields  $s = 2$ ,  $t = 2$ ,  $u = 3$ . Putting  $u = 3$  into the line  $q$  equation, gives  $Q = [5, 5, 3]$ . The desired transversal is thus given by  $Q$  and the point  $[1, 0, 0]$ . The intersection of the transversal with  $p$  is at  $P = [7/3, 5/3, 1]$ .  $\square$

**4.A.16.** Find the common perpendicular between the two skew lines

$$\begin{aligned} p: & [3, 0, 3] + (0, 1, 2)t, \quad t \in \mathbb{R} \\ q: & [0, -1, -2] + (1, 2, 3)s \quad s \in \mathbb{R} \end{aligned}$$

**Solution.** The direction of the common perpendicular is given by the cross product of the two direction vectors. So the direction of the common perpendicular is  $(1, -2, 1)$ . Form a linear equation system which

Examples of convex sets are

- (1) the empty set,
- (2) affine subspaces,
- (3) line segments, rays  $p = \{P + t \cdot v; t \geq 0\}$ ,
- (4) more generally  $k$ -dimensional objects

$$\alpha = \{P + t_1 \cdot v_1 + \dots + t_k \cdot v_k; t_1, \dots, t_k \in \mathbb{R}, t_k \geq 0\},$$

- (5) angles in two-dimensional subspaces

$$\beta = \{P + t_1 \cdot v_1 + t_2 \cdot v_2; t_1 \geq 0, t_2 \geq 0\}.$$

The intersection of an arbitrary system of convex sets is a convex set. The intersection of all convex sets containing a given set  $M$  is called the *convex hull*  $\mathcal{K}(M)$  of the set  $M$ .

**Theorem.** The convex hull of any subset  $M \subset \mathcal{A}$  is

$$\mathcal{K}(M) = \{t_1 A_1 + \dots + t_s A_s; \sum_{i=1}^s t_i = 1, t_i \geq 0, A_i \in M\}$$



**PROOF.** Let  $S$  denote the set of all affine combinations on the right-hand side of the equation. To check that  $S$  is convex, choose two sets of parameters  $t_i, i = 1, \dots, s_1, t'_j, j = 1, \dots, s_2$  with the desired properties.

Without loss of generality, assume that  $s_1 = s_2$  and that the same points from  $M$  there appear in both combinations (otherwise simply add summands with zero coefficients). Consider an arbitrary point on the line segment given by the vertices defined by the two combinations:

$$\varepsilon(t_1 A_1 + \dots + t_s A_s) + (1 - \varepsilon)(t'_1 A_1 + \dots + t'_s A_s), \quad 0 \leq \varepsilon \leq 1.$$

Obviously any point of this line segment lies in  $S$ .

It remains to show that the convex hull of the points  $A_1, \dots, A_s$  cannot be smaller than  $S$ . The points  $A_i$  themselves correspond to the choice of parameters  $t_j = 0$  for all  $j \neq i$  and  $t_i = 1$ . Assume that the claim holds for all sets with at most  $s - 1$  points. Then the convex hull of the points  $A_1, \dots, A_{s-1}$  is (according to the assumption) formed exactly by the combinations from the right side of the equation to be proved, where  $t_s = 0$ . Now consider a point  $A = t_1 A_1 + \dots + t_s A_s \in S, t_s < 1$ , and affine combinations  $\varepsilon(t_1 A_1 + \dots + t_{s-1} A_{s-1}) + (1 - \varepsilon(1 - t_s)) A_s, 0 \leq \varepsilon \leq \frac{1}{1 - t_s}$ . It is a line segment with vertices given by parameters  $\varepsilon = 0$  (the point  $A_s$ ) and  $\varepsilon = 1/(1 - t_s)$  (a point in the convex hull of  $A_1, \dots, A_{s-1}$ ). The point  $A$  is an inner point of this line segment with the parameter  $\varepsilon = 1$ , and thus  $A$  lies in the convex hull of  $A_1, \dots, A_s$ .  $\square$

The convex hulls of finite sets are called *convex polyhedrons*.

We have a  $k$ -dimensional *simplex* if and only if the vertices  $A_0, \dots, A_k$  defining the convex polyhedron are in general position. In the case of a simplex, the expression of any of its points as an affine combination of the defining vertices is unique.

Specific examples are the convex polyhedrons defined by one point and a finite number of vectors. Let  $u_1, \dots, u_k$  be

expresses that a vector defined by two points, one lying on  $p$ , the other on  $q$ , is parallel to the direction  $(1, -2, 1)$ . We get the system  $P - Q = k(1, -2, 1)$ , or  $\underbrace{[3, 0, 3] + (0, 1, 2)t}_{P} - \underbrace{[0, -1, -2] + (1, 2, 3)s}_{Q} = k(1, -2, 1)$ .

Treat this equality component-wise to give

$$\begin{aligned} 3 - s &= k \\ 1 + t - 2s &= -2k \\ 5 + 2t - 3s &= k \end{aligned}$$

with the solution  $t = 1, s = 2, k = 1$ . Put  $t = 1$  into the line  $p$ , to obtain the point  $[3, 1, 5]$  on the common perpendicular. Put  $s = 2$  into the line  $q$  equation to obtain the point  $[3, 1, 5]$ . The common perpendicular is defined by the line joining these two points.  $\square$

### B. Euclidean geometry

**4.B.1.** Determine the distance between the lines in  $\mathbb{R}^3$ .

$p : [1, -1, 0] + t(-1, 2, 3)$ , and  $q : [2, 5, -1] + t(-1, -2, 1)$ .

**Solution.** The distance is defined as the distance of the orthogonal projection of arbitrary points on the respective lines to the orthogonal complement of the vector subspace generated by their directions. The orthogonal complement is spanned by the cross product:

$$\begin{aligned} \langle (-1, 2, 3), (-1, -2, 1) \rangle^\perp &= \langle (-1, 2, 3) \times (-1, -2, 1) \rangle \\ &= \langle (8, -2, 4) \rangle = \langle (4, -1, 2) \rangle. \end{aligned}$$

A transversal is (for example) the segment joining  $[1, -1, 0]$  to  $[2, 5, -1]$ . So the vector to be projected is  $[1, -1, 0] - [2, 5, -1] = (-1, -6, 1)$ . The distance between the lines is therefore:

$$\rho(p, q) = \frac{|(-1, -6, 1) \cdot (4, -1, 2)|}{\|(4, -1, 2)\|} = \frac{4}{\sqrt{21}}.$$

$\square$

**4.B.2.** Find a point  $A$  lying on the line

$$p : x + 2y + z - 1 = 0, \quad 3x - y + 4z - 29 = 0,$$

which is equidistant from both  $B = [3, 11, 4]$  and  $C = [-5, -13, -2]$ .

**Solution.** First, express the line  $p$  parametrically. Solve the system

$$\begin{aligned} x + 2y + z &= 1, \\ 3x - y + 4z &= 29. \end{aligned}$$

Rewrite the system as an augmented matrix and perform row operations

$$\begin{aligned} \left( \begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 3 & -1 & 4 & 29 \end{array} \right) &\sim \left( \begin{array}{ccc|c} 1 & 2 & 1 & 1 \\ 0 & -7 & 1 & 26 \end{array} \right) \\ &\sim \left( \begin{array}{ccc|c} 1 & 0 & 9/7 & 59/7 \\ 0 & 1 & -1/7 & -26/7 \end{array} \right). \end{aligned}$$

The line  $p$  is thus described by

$$p : \left[ \frac{59}{7}, -\frac{26}{7}, 0 \right] + t \left( -\frac{9}{7}, \frac{1}{7}, 1 \right), \quad t \in \mathbb{R}.$$

arbitrary vectors in the difference space  $\mathbb{R}^n$ ,  $A \in \mathcal{A}_n$  a point. A *parallelepiped*  $\mathcal{P}_k(A; u_1, \dots, u_k) \subset \mathcal{A}_n$  is the set

$$\mathcal{P}_k(A; u_1, \dots, u_k) = \{A + c_1u_1 + \dots + c_ku_k; 0 \leq c_i \leq 1\}.$$

If the vectors  $u_1, \dots, u_k$  are independent, we talk about a  $k$ -dimensional parallelepiped  $\mathcal{P}_k(A; u_1, \dots, u_k) \subset \mathcal{A}_n$ . It is clear from the definition that parallelepipeds are convex. They are the convex hulls of their vertices.

**4.1.10. Examples of standard affine tasks.** (1) *To find a parametric description of an implicitly given subspace and vice versa:*



Find a particular solution of a non-homogeneous system and a fundamental solution of the homogenized system. Then obtain (in the coordinates in which the equations have been set) the desired parametric description. In the opposite direction, write the parametric description in coordinates and then eliminate the free parameters  $t_1, \dots, t_k$ . This results in the equations defining the given subspace implicitly.

(2) *To find the subspace generated by several subspaces  $\mathcal{Q}_1, \dots, \mathcal{Q}_s$  (of different dimensions in general). For instance, to find a plane in  $\mathcal{A}_3$  given by a straight line and a point, or by three points). To define this subspace implicitly or parametrically:*

The resulting subspace  $\mathcal{Q}$  is always determined by one fixed point  $A_i$  in each subspace  $\mathcal{Q}_i$  and by the sum of all difference spaces. For instance,

$$\mathcal{Q} = A_1 + (Z(\{A_1, \dots, A_s\}) + Z(\mathcal{Q}_1) + \dots + Z(\mathcal{Q}_s)).$$

If the subspaces are given implicitly, it is possible to convert them into parametric form first. Nevertheless, different methods are advantageous in some concrete situations. Notice that it is really necessary to use one point from each of the subspaces. For example, two parallel lines in a plane generate the whole plane, but they share the same one-dimensional difference space.

(3) *To find the intersection of the subspaces  $\mathcal{Q}_1, \dots, \mathcal{Q}_s$ :*

If they are given in the implicit form, it is sufficient to unify all equations into one system, omitting any linearly dependent ones. If the resulting system has no solution, then the intersection is empty. Otherwise, an implicit description of the affine subspace is obtained. This is the intersection we are searching for.

If parametric forms are given, we may search directly for common points as solutions of the appropriate equations, similarly to the way we find the intersections of vector spaces. If the number of subspaces is greater than two, we must search for the intersection step by step.

If one of the subspaces is defined parametrically and the other implicitly, it suffices to substitute the parametrized coordinates and to solve the resulting system of equations.

(4) *To find a crossbar between two skew lines  $p, q$  in  $\mathcal{A}_3$  passing through a given point or having a given direction:*

By a *crossbar* we mean a straight line which has nonempty intersection with both skew lines. Thus the

It is convenient to avoid the fractions, by introducing the substitution  $t = 7s + 26$ .  $p$  is thus described by

$$p : [-25, 0, 26] + s(-9, 1, 7), \quad s \in \mathbb{R}.$$

The point  $A$  is obtained by requiring that the vectors

$$A - B = (-28 - 9s, -11 + s, 22 + 7s),$$

$$A - C = (-20 - 9s, 13 + s, 28 + 7s)$$

have the same length. Hence

$$\begin{aligned} & \sqrt{(-28 - 9s)^2 + (-11 + s)^2 + (22 + 7s)^2} \\ &= \sqrt{(-20 - 9s)^2 + (13 + s)^2 + (28 + 7s)^2}, \end{aligned}$$

or rather

$$\begin{aligned} & (-28 - 9s)^2 + (-11 + s)^2 + (22 + 7s)^2 \\ &= (-20 - 9s)^2 + (13 + s)^2 + (28 + 7s)^2. \end{aligned}$$

which has the unique solution  $s = -3$ . Therefore

$$A = [-25, 0, 26] - 3(-9, 1, 7) = [2, -3, 5].$$

□

**4.B.3.** Michael has a stick of length 4. Can he touch the lines  $p$  and  $q$  simultaneously with this stick, given that the stick must pass through  $[2, 1, 2]$ ?

$$\begin{aligned} p & : [-1, 4, 1] + t(-1, 2, 0), \\ q & : [4, 4, -1] + s(1, 2, -4)? \end{aligned}$$

**Solution.** Compute the transversal of those lines passing through  $[2, 1, 2]$ . It is the segment joining  $[1, 0, 1]$  to  $[3, 2, 3]$ . Its length is  $\sqrt{12}$ , which is less than 4. So Michael can touch the lines as required. □

**4.B.4.** In Euclidean space  $\mathbb{R}^4$ , determine the distance between the point  $A = [2, -5, 1, 4]$  and the subspace defined by the equations

$$\begin{aligned} U : 4x_1 - 2x_2 - 3x_3 - 2x_4 + 12 &= 0, \\ 2x_1 - x_2 - 2x_3 - 2x_4 + 9 &= 0. \end{aligned}$$

**Solution.** Find first a parametric expression of the subspace  $U$ . For example,

$$B = [0, 3, 0, 3] \in U.$$

The distance between  $A$  and  $U$  equals the length of the orthogonal projection of the vector  $A - B$  to the orthogonal complement of the direction of the subspace  $U$ . However, the orthogonal complement of the  $U$  direction (it defines this subspace) – as set (of linear combination of normal vectors)

$$V := \{t(4, -2, -3, -2) + s(2, -1, -2, -2); t, s \in \mathbb{R}\}.$$

We need to find the orthogonal projection  $P_{A-B}$  of vector  $A - B$  to  $V$ , which lies in  $V$ , and thus

$$P_{A-B} = a(4, -2, -3, -2) + b(2, -1, -2, -2)$$

for certain  $a, b \in \mathbb{R}$ . Clearly,  $(A - B - P_{A-B}) \perp V$ , thus

$$((A - B) - P_{A-B}) \perp (4, -2, -3, -2),$$

resulting crossbar  $r$  is a one-dimensional affine subspace. If we are given one point  $A \in r$ , then the affine subspace generated by  $p$  and  $A$  is either a straight line (if  $A \in p$ ) or a plane (if  $A \notin p$ ). In the first case, there are an infinite number of solutions, one for each point of  $q$ . In the second case, it suffices to find the intersection  $B$  of the plane  $\langle p \cup A \rangle$  with  $q$ , and  $r = \langle \{A, B\} \rangle$ . There is no solution if the intersection is empty. If  $q \subset \langle p \cup A \rangle$ , there are an infinite number of solutions. If the intersection has one element, there is exactly one solution.

If a direction  $u \in \mathbb{R}^n$  is given, then we consider the subspace  $Q$  generated by  $p$  and the difference space  $Z(p) + \langle u \rangle \subset \mathbb{R}^n$ . Again, we obtain an infinite number of solutions if  $q \subset Q$ . Otherwise we consider the intersection  $Q$  with  $q$  and we finish as before.

The solutions of other practical geometric problems are based mostly on the systematic use of the steps given above.

**4.1.11. Remark on linear programming.** In the beginning of the third chapter in paragraphs 3.1.1–3.1.7, we dealt with practical problems which are given by systems of linear inequalities. Each single inequality



$$a_1x_1 + \cdots + a_nx_n \leq b$$

defines a halfspace in the standard affine space  $\mathbb{R}^n$ . This is bounded by a hyperplane given by the corresponding equation (compare with the definition in paragraph 4.1.9(4)).

In particular, the set of all admissible vectors for the problem of the linear programming is always an intersection of a finite number of convex sets. Hence the set itself is either convex or empty.

If the intersection is both nonempty and bounded, then it is a convex polyhedron. As justified in 3.1.1 already, each linear form is either increasing or decreasing or constant along each parametrized straight line in the affine space. Thus if a given problem from linear programming is solvable and bounded, then it has the optimal solution at one of the vertices of the corresponding convex polyhedron. The reader should be able to imagine this claim in the case of two-dimensional or three-dimensional problems. Nevertheless, the straightforward explanation in these low dimensions holds for all finite-dimensional cases.

We have given a “geometric proof” of the existence part of the fundamental theorem 3.1.5. We have translated the initial problem into a finite problem of the given cost function.

**4.1.12. Euclidean point spaces.** So far, we do not need the notions of distance and length for geometric considerations. But the length of vectors and the angle between vectors, as defined in the second chapter (see 2.3.18 and elsewhere), play a significant role in many practical problems.



$$((A - B) - P_{A-B}) \perp (2, -1, -2, -2).$$

By substitution of  $A - B$  and  $P_{A-B}$ ,

$$((2, -8, 1, 1) - a(4, -2, -3, -2) - b(2, -1, -2, -2)) \cdot (4, -2, -3, -2) = 0,$$

$$((2, -8, 1, 1) - a(4, -2, -3, -2) - b(2, -1, -2, -2)) \cdot (2, -1, -2, -2) = 0;$$

so

$$\begin{aligned} &(2, -8, 1, 1) \cdot (4, -2, -3, -2) \\ &- a(4, -2, -3, -2) \cdot (4, -2, -3, -2) \\ &- b(2, -1, -2, -2) \cdot (4, -2, -3, -2) = 0, \end{aligned}$$

$$\begin{aligned} &((2, -8, 1, 1) \cdot (2, -1, -2, -2)) \\ &- a(4, -2, -3, -2) \cdot (2, -1, -2, -2) \\ &- b(2, -1, -2, -2) \cdot (2, -1, -2, -2) = 0. \end{aligned}$$

If we compute these dot products, we obtain the system

$$\begin{aligned} 19 - 33a - 20b &= 0, \\ 8 - 20a - 13b &= 0, \end{aligned}$$

with the only solution  $a = 3, b = -4$ . Hence

$$\begin{aligned} P_{A-B} &= 3(4, -2, -3, -2) - 4(2, -1, -2, -2) \\ &= (4, -2, -1, 2), \end{aligned}$$

where

$$\|P_{A-B}\| = \sqrt{4^2 + (-2)^2 + (-1)^2 + 2^2} = 5.$$

Hence the distance between  $A$  and  $U$  equals  $\|P_{A-B}\| = 5$ .

□

**4.B.5.** In the vector space  $\mathbb{R}^4$ , compute the distance  $v$  between the point  $[0, 0, 6, 0]$  and the vector subspace

$$U : [0, 0, 0, 0] + t_1(1, 0, 1, 1) + t_2(2, 1, 1, 0) + t_3(1, -1, 2, 3), \\ t_1, t_2, t_3 \in \mathbb{R}$$

**Solution.** We solve the problem by the least squares method. Write the generating vectors of  $U$  as the columns of the matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix}.$$

Substitute the point  $[0, 0, 6, 0]$  by the corresponding vector  $b = (0, 0, 6, 0)^T$ . Now solve  $A \cdot x = b$ . This is the linear equation system

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 0, \\ &x_2 - x_3 = 0, \\ x_1 + x_2 + 2x_3 &= 6, \\ x_1 + 3x_3 &= 0, \end{aligned}$$

by the least squares method. (Note that the system does not have a solution – the distance would be 0 otherwise.) Multiply

EUCLIDEAN SPACES

The *standard Euclidean point space*  $\mathcal{E}_n$  is the affine space  $\mathcal{A}_n$  whose difference space is the standard *Euclidean space*  $\mathbb{R}^n$  with the scalar product

$$\langle x, y \rangle = y^T \cdot x.$$

The *Cartesian coordinate system* is the affine coordinate system  $(A_0; \underline{u})$  with the orthonormal basis  $\underline{u}$ .

The *Euclidean distance* between two points  $A, B \in \mathcal{E}_n$  is defined as the length of the vector  $\|B - A\|$ . This is denoted by  $\rho(A, B)$ .

*Euclidean subspaces* in  $\mathcal{E}_n$  are affine subspaces, where the corresponding difference spaces are considered with restricted scalar products.

By a *Euclidean point space*  $\mathcal{E}$  of dimension  $n$  is meant an affine space, whose difference space is a real  $n$ -dimensional Euclidean vector space. The notion of a Cartesian coordinate system has an obvious meaning. Since each choice of such a coordinate system identifies  $\mathcal{E}$  with the standard space  $\mathcal{E}_n$ , we deal with the standard Euclidean spaces and their subspaces, with no loss of generality.

From the geometric point of view, simple properties of the scalar product like the triangular inequality, the Cauchy inequality, Bessel's inequality, derived in the previous chapter (see 3.4.3), have useful consequences:



**4.1.13. Theorem.** For points  $A, B, C \in \mathcal{E}_n$  the following holds

- (1)  $\rho(A, B) = \rho(B, A)$
- (2)  $\rho(A, B) = 0$  if and only if  $A = B$
- (3)  $\rho(A, B) + \rho(B, C) \geq \rho(A, C)$
- (4) In each Cartesian coordinate system  $(A_0; \underline{e})$ , the distance between the points  $A = A_0 + a_1e_1 + \dots + a_n e_n$ ,  $B = A_0 + b_1e_1 + \dots + b_n e_n$  is  $\sqrt{\sum_{i=1}^n (a_i - b_i)^2}$ .
- (5) Given a point  $A$  and a subspace  $\mathcal{Q}$  in  $\mathcal{E}_n$ , there exists a point  $P \in \mathcal{Q}$  which minimizes the distance between  $A$  and points in  $\mathcal{Q}$ . The distance between  $A$  and  $P$  equals the length of the orthogonal projection of the vector  $A - B$  into  $Z(\mathcal{Q})^\perp$  for an arbitrary  $B \in \mathcal{Q}$ .
- (6) More generally, for subspaces  $\mathcal{Q}$  and  $\mathcal{R}$  in  $\mathcal{E}_n$  there exist points  $P \in \mathcal{Q}$  and  $Q \in \mathcal{R}$  which minimize the distance between points  $B \in \mathcal{Q}$  and  $A \in \mathcal{R}$ . The distance between the points  $P$  and  $Q$  is the length of the orthogonal projection of the vector  $A - B$  into  $(Z(\mathcal{Q}) + Z(\mathcal{R}))^\perp$  for arbitrary points  $B \in \mathcal{Q}$  and  $A \in \mathcal{R}$ .

**PROOF.** The first three properties follow directly from the properties of the length of vectors in spaces with a scalar product. The fourth follows from the expression of the scalar product in an orthonormal basis.

Consider the relation for the minimal distances  $\rho(A, B)$  for  $B \in \mathcal{Q}$ . The vector  $A - B$  decomposes uniquely as  $A - B = u_1 + u_2$ , where  $u_1 \in Z(\mathcal{Q})$ ,  $u_2 \in Z(\mathcal{Q})^\perp$ . The component  $u_2$  does not depend on the choice of  $B \in \mathcal{Q}$ . This

$A \cdot x = b$  by the matrix  $A^T$  from the left-hand side. Then the augmented matrix  $A^T \cdot A \cdot x = A^T \cdot b$  is

$$\left( \begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 3 & 6 & 3 & 6 \\ 6 & 3 & 15 & 12 \end{array} \right).$$

By elementary row operations, transform the matrix to the normal form

$$\begin{aligned} \left( \begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 3 & 6 & 3 & 6 \\ 6 & 3 & 15 & 12 \end{array} \right) &\sim \left( \begin{array}{ccc|c} 3 & 3 & 6 & 6 \\ 0 & 3 & -3 & 0 \\ 0 & -3 & 3 & 0 \end{array} \right) \\ &\sim \left( \begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right). \end{aligned}$$

Continue with backward elimination

$$\left( \begin{array}{ccc|c} 1 & 1 & 2 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \sim \left( \begin{array}{ccc|c} 1 & 0 & 3 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right).$$

The solution is

$$x = (2 - 3t, t, t)^T, \quad t \in \mathbb{R}.$$

Note that the existence of infinitely many solutions is caused by third vector generating  $U$ , which is redundant because

$$3(1, 0, 1, 1) - (2, 1, 1, 0) = (1, -1, 2, 3).$$

An arbitrary ( $t \in \mathbb{R}$ ) linear combination

$$(2 - 3t)(1, 0, 1, 1) + t(2, 1, 1, 0) + t(1, -1, 2, 3) = (2, 0, 2, 2)$$

corresponds to a point  $[2, 0, 2, 2]$  in the subspace  $U$ , which is the nearest point to  $[0, 0, 6, 0]$ . The required distance is therefore

$$\begin{aligned} v &= \|[2, 0, 2, 2] - [0, 0, 6, 0]\| = \sqrt{2^2 + 0 + (-4)^2 + 2^2} \\ &= 2\sqrt{6}. \end{aligned}$$

□

**4.B.6.** Compute the volume of the parallelepiped in  $\mathbb{R}^3$  with base in the plane  $z = 0$  and with edges given by pairs of vertices  $[0, 0, 0]$ ,  $[-2, 3, 0]$ ;  $[0, 0, 0]$ ,  $[4, 1, 0]$  and  $[0, 0, 0]$ ,  $[5, 7, 3]$ .

**Solution.** The parallelepiped is given by vectors  $(4, 1, 0)$ ,  $(-2, 3, 0)$ ,  $(5, 7, 3)$ . Its volume is the determinant

$$\begin{vmatrix} 4 & -2 & 5 \\ 1 & 3 & 7 \\ 0 & 0 & 3 \end{vmatrix} = 3 \begin{vmatrix} 4 & -2 \\ 1 & 3 \end{vmatrix} = 3 \cdot 14 = 42.$$

Note that if the order of vectors is changed, we would get result  $\pm 42$ , because the determinant gives the *oriented* volume of parallelepiped. Note further that the volume would not change if the third vector was  $[a, b, 3]$  for arbitrary  $a, b \in \mathbb{R}$ . Its surface depends only on orthogonal distance between planes of its upper and lower base and their area

$$\begin{vmatrix} 4 & -2 \\ 1 & 3 \end{vmatrix} = 14.$$

□

is because any potential change of  $B$  is apparent by adding a vector from  $Z(Q)$ .

Choose  $P = A + (-u_2) = B + u_1 \in Q$ . Then

$$\|A - B\|^2 = \|u_1\|^2 + \|u_2\|^2 \geq \|u_2\|^2 = \|A - P\|^2.$$

Hence the minimal distance is obtained for the point  $P$ . Its value is  $\|u_2\|$ .

The general result is obtained in a similar way. For the choice of arbitrary points  $A \in \mathcal{R}$  and  $B \in \mathcal{Q}$  their difference is given as a sum of vectors  $u_1 \in Z(\mathcal{R}) + Z(\mathcal{Q})$  and  $u_2 \in (Z(\mathcal{R}) + Z(\mathcal{Q}))^\perp$ . The component  $u_2$  does not depend on the choice of the points. By adding suitable vectors from the difference spaces of  $\mathcal{R}$  and  $\mathcal{Q}$ , points  $A'$  and  $B'$  are obtained so that the distance between them is  $\|u_2\|$ . □

We consider more elementary problems in affine geometry requiring the concept of distance.

**4.1.14. Examples of standard problems.** (1) To find the distance from the point  $A \in \mathcal{E}_n$  to the subspace  $Q \subset \mathcal{E}_n$ :



A method of solving such a problem is given in proposition 4.1.13.

(2) In  $\mathcal{E}_2$  to construct the straight line  $q$  through a given point  $A$  which forms a given angle with a given line  $p$ :

Recall that we work with angles between vectors in plane geometry already (see e.g. 2.3.22). Find a vector  $u \in \mathbb{R}^2$  lying in the difference space of the line  $p$ . Then choose a vector  $v$  having the prescribed angle with  $u$ . The desired line is given by the point  $A$  and the difference space  $\langle v \rangle$ . The problem has either one or two solutions.

(3) To find a line through a given point, perpendicular to a given line:

The procedure is introduced in the proof of the last but one item of proposition 4.1.13.

(4) In  $\mathcal{E}_3$  to determine the distance between two lines  $p, q$ :

Choose any point from each of the lines,  $A \in p, B \in q$ . The component of the vector  $A - B$  lying in the orthogonal complement  $(Z(p) + Z(q))^\perp$  has length equal to the distance between  $p$  and  $q$ .

(5) In  $\mathcal{E}_3$  to find the axis of two skew lines  $p$  a  $q$ :

By the axis we mean the crossbar which attains the minimal possible distance between the given skew lines in terms of the points of intersection. The procedure can be derived from the proof of proposition 4.1.13 (the last item). Let  $\eta$  be the subspace generated by a single point  $A \in p$  and the sum  $Z(p) + (Z(p) + Z(q))^\perp$ . Provided that the lines  $p$  and  $q$  are not parallel,  $\eta$  is a plane. Then the intersection  $\eta \cap q$  together with the difference space  $(Z(p) + Z(q))^\perp$  gives a parametric description of the desired axis. If the lines are parallel, the problem has an infinite number of solutions.

**4.B.7.** Let the points  $[0, 0, 1]$ ,  $[2, 1, 1]$ ,  $[3, 3, 1]$ ,  $[1, 2, 1]$  define a parallelogram. Determine the point  $X$  lying on the line  $p : [0, 0, 1] + (1, 1, 1)t$  so that the parallelepiped defined by the given parallelogram and a point  $X$  has volume of 1.

**Solution.** Form a determinant which gives the volume of a parallelepiped with  $X$  moving along line  $p$ :

$$\begin{vmatrix} t & t & t \\ 2 & 1 & 0 \\ 1 & 2 & 0 \end{vmatrix}.$$

The volume is  $3t$  which implies  $t = 1/3$ . □

**4.B.8.** Let  $ABCDEFGH$  be a cube (with common notation, i.e. vectors  $E - A$ ,  $F - B$ ,  $G - C$ ,  $H - D$  are orthogonal to the plane defined by vertices  $A, B, C, D$ ) in Euclidean space  $\mathbb{R}^3$ . Compute the angle  $\varphi$  between the vectors  $F - A$  and  $H - A$ .

**Solution.** This problem is solved using the formula for the angle between the vectors. Alternatively notice that the vertices  $A, F, H$  are the vertices of a triangle with all sides of the same length. Hence it is an equilateral triangle. Therefore  $\varphi = \pi/3$ . □

**4.B.9.** Let  $S$  be the midpoint of the edge  $AB$  of the cube  $ABCDEFGH$  (with common labelling). Compute the cosine of the angle between the lines  $ES$  and  $BG$ .

**Solution.** Dilation (homothety) is a mapping which preserves angles. So without loss of generality, the cube edge has length 1. The coordinate system can be placed so that  $A$  is at the origin,  $B = [1, 0, 0]$  and  $E = [0, 0, 1]$ . It follows that then:  $S = [1/2, 0, 0]$ ,  $G = [1, 1, 1]$ ,  $ES = (1/2, 0, -1)$  and  $BG = (0, 1, 1)$ . The desired cosine of the angle  $\varphi$  is then

$$\cos(\varphi) = \left| \frac{(1/2, 0, -1) \cdot (0, 1, 1)}{\|(1/2, 0, -1)\| \|(0, 1, 1)\|} \right| = \frac{\sqrt{2}}{\sqrt{5}}$$

**4.B.10.** Compute the angle between the line  $p$  given by the implicit equations

$$\begin{aligned} x + 3y + z &= 0, \\ -x - y + z &= 0 \end{aligned}$$

and plane the  $\varrho : x + y + 2z + 1 = 0$ .

**Solution.** The normal vector of the plane  $\varrho$  is  $(1, 1, 2)$ . Copy the first equation of the line  $p$ . Sum both of them, to obtain

$$\begin{aligned} x + 3y + z &= 0, \\ 2y + 2z &= 0. \end{aligned}$$

From this system  $y = -z$  and  $x = 2z$ . The vector  $(2, -1, 1)$  is therefore the direction vector of  $p$ . In other words,  $p$  passes through the origin, and

$$p : [0, 0, 0] + t(2, -1, 1), \quad t \in \mathbb{R}.$$

For the angle  $\varphi$  between the vectors  $(1, 1, 2)$ ,  $(2, -1, 1)$ ,

$$\cos \varphi = \frac{2 - 1 + 2}{\sqrt{6} \cdot \sqrt{6}} = \frac{1}{2}.$$

**4.1.15. Angles.** Various geometric notions like angles, orientation, volume etc. in the point spaces  $\mathcal{E}_n$  are defined in terms of suitable notions from Euclidean spaces. The angle between two vectors is defined at the end of the third part of the second chapter, see 2.3.22.



From the Cauchy inequality, it follows that  $0 \leq \frac{|u \cdot v|}{\|u\| \|v\|} \leq 1$ . So it makes sense to define the *angle*  $\varphi(u, v)$  between vectors  $u, v \in V$  in a real vector space with a scalar product given by the equation

$$\cos \varphi(u, v) = \frac{u \cdot v}{\|u\| \|v\|}, \quad 0 \leq \varphi(u, v) \leq 2\pi.$$

This is completely in accordance with the situation in two-dimensional Euclidean space  $\mathbb{R}^2$  and with the philosophy that the notion related to the two vectors is the issue of plane geometry.

In the Euclidean plane, we use also the geometric functions  $\cos$  and  $\sin$  defined by a pure geometric consideration. Therefore, the angle between two vectors in higher-dimensional spaces is measured in the plane which is generated by these two vectors (or it is zero).

In an arbitrary real vector space with a scalar product, it follows that

$$\begin{aligned} \|u - v\|^2 &= \|u\|^2 + \|v\|^2 - 2(u \cdot v) \\ &= \|u\|^2 + \|v\|^2 - 2\|u\| \|v\| \cos \varphi(u, v). \end{aligned}$$

This is the well known *cosine rule* from plane geometry.

Consider an orthonormal basis  $\underline{e}$  of the difference space  $V$  and a vector  $u \in V$ . The square of the size of  $u$  is given by the usual formula

$$\|u\|^2 = \sum_i |u \cdot e_i|^2.$$

By dividing this equation by the number  $\|u\|^2$ ,

$$1 = \sum_i (\cos \varphi(u, e_i))^2,$$

which is known as the *law of direction cosines*  $\cos \varphi(u, e_i)$  of the vector  $u$ .

Now we can derive definitions for angles between general subspaces in a Euclidean vector space from the definitions of angles between vectors. In particular, it must be decided how to deal with cases where the subspaces have a nontrivial intersection. For the angle between two lines, use the smaller of the two possible angles.

In the case of two nonparallel planes in  $\mathbb{R}^3$  we do not say that the angle is zero. They intersect and have one direction in common. Perhaps we take the two perpendicular lines to this common direction in the two planes and measure their angle.

The general cases are treated as follows:





Hence  $\varphi = 60^\circ$ . However, this is the angle between the direction vector of  $p$  and the normal vector  $\rho$ . The desired angle is the complement of this angle, so the solution is  $30^\circ = 90^\circ - 60^\circ$ .  $\square$

**4.B.11.** In the real plane, find a line which passes through the point  $[-3, 0]$ , so that  $60^\circ$  is the angle between this line and the line

$$p : \sqrt{3}x + 3y + 5 = 0.$$

**Solution.** The given line has slope  $-\frac{1}{\sqrt{3}}$ . This is at angle  $-30^\circ$  from the positive  $x$  axis. Thus the required line is either at angle  $-90^\circ$  or angle  $30^\circ$  from the positive  $x$  axis. The former determines the vertical line  $x = -3$ . The latter determines the line with slope  $\frac{1}{\sqrt{3}}$  through  $[-3, 0]$ , hence has equation  $y\sqrt{3} = x + 3$ .  $\square$

**Solution.** (Alternative) Notice that there are two such lines. The general equation of a line in the plane has the form

$$ax + by + c = 0. \quad \text{Choose parameters so that } a^2 + b^2 = 1.$$

We find such numbers  $a, b, c \in \mathbb{R}$ , so that all the conditions are satisfied. Since the line passes through  $[-3, 0]$ ,  $c = 3a$ . The condition of the angle between lines equals  $60^\circ$  then gives

$$\frac{1}{2} = \cos 60^\circ = \frac{|\sqrt{3}a + 3b|}{\sqrt{12}}, \quad \text{tj. } \sqrt{3} = |\sqrt{3}a + 3b|.$$

Performing further operations

$$\pm 1 = a + \sqrt{3}b \quad \text{and exponentiation} \quad 1 = a^2 + 3b^2 + 2\sqrt{3}ab.$$

If we use  $a^2 + b^2 = 1$ , we get

$$0 = 2b^2 + 2\sqrt{3}ab, \quad \text{tj. } 0 = b(b + \sqrt{3}a).$$

Together (remember that  $c = 3a$  and  $a^2 + b^2 = 1$ )

$$a = \pm 1, \quad b = 0, \quad c = \pm 3; \quad a = \pm \frac{1}{2}, \quad b = \mp \frac{\sqrt{3}}{2}, \quad c = \pm \frac{3}{2}$$

We can easily check that lines determined by those coefficients

$$x + 3 = 0, \quad \frac{1}{2}x - \frac{\sqrt{3}}{2}y + \frac{3}{2} = 0$$

satisfy all the conditions.  $\square$

**4.B.12.** Determine the equation of all planes so that the angle between every such plane and the plane  $x + y + z - 1 = 0$  is  $60^\circ$ , and further, that they contain the line  $p : [1, 0, 0] + t(1, 1, 0)$ .  $\circ$

**4.B.13.** Determine the angle between the planes

$$\sigma : [1, 0, 2] + (1, -1, 1)t + (0, 1, -2)s$$

$$\rho : [3, 3, 3] + (1, -2, 0)t + (0, 1, 1)s$$

**Solution.** The line of intersection between the planes has direction vector  $(1, -1, 1)$ . The plane orthogonal to this vector has intersection with the given planes generated by the vectors  $(1, 0, -1)$  and  $(0, 1, 1)$ . The angle between these one-dimensional subspaces is  $60^\circ$ .  $\square$

ANGLES BETWEEN SUBSPACES

**4.1.16. Definition.** Consider finite-dimensional subspaces  $U_1, U_2$  in a Euclidean vector space  $V$  of arbitrary dimension.

The angle between vector subspaces  $U_1, U_2$  is the real number  $\alpha = \varphi(U_1, U_2) \in [0, \frac{\pi}{2}]$  satisfying:

(1) If  $\dim U_1 = \dim U_2 = 1, U_1 = \langle u \rangle, U_2 = \langle v \rangle$ , then

$$\cos \alpha = \frac{|u \cdot v|}{\|u\| \|v\|}.$$

(2) If the dimensions of  $U_1, U_2$  are positive, and if  $U_1 \cap U_2 = \{0\}$ , then the angle is the minimum of all angles between the one-dimensional subspaces

$$\alpha = \min\{\varphi(\langle u \rangle, \langle v \rangle); 0 \neq u \in U_1, 0 \neq v \in U_2\}.$$

Such a minimum always exists.

(3) If  $U_1 \subset U_2$  or  $U_2 \subset U_1$  (in particular if one of them is empty), then  $\alpha = 0$ .

(4) If  $U_1 \cap U_2 \neq \{0\}$  and if  $U_1 \neq U_1 \cap U_2 \neq U_2$ , then

$$\alpha = \varphi(U_1 \cap (U_1 \cap U_2)^\perp, U_2 \cap (U_1 \cap U_2)^\perp).$$

The angle between affine subspaces  $\mathcal{Q}_1, \mathcal{Q}_2$  in a Euclidean point space  $\mathcal{E}_n$  is defined as the angle between their difference spaces  $Z(\mathcal{Q}_1), Z(\mathcal{Q}_2)$ .

Notice that the angle is always well defined. In the last case,

$$(U_1 \cap (U_1 \cap U_2)^\perp) \cap (U_2 \cap (U_1 \cap U_2)^\perp) = \{0\}$$

so we can determine the angle according to item (2). Notice also that in the case  $U_1 \cap U_2 = \{0\}$ , the subspaces  $U_1$  and  $U_2$  are perpendicular in terms of the former definitions if and only if the angle between them is  $\pi/2$ . However, if the intersection is nontrivial, then they cannot be perpendicular in the former sense.

In order to show the validity of the definition, it remains to show that the vectors  $u \in U_1, v \in U_2$  minimizing the expression for the angle always exist. First a special case:

**4.1.17. Lemma.** Let  $v$  be a vector in a Euclidean space  $V$  and  $U \subset V$  an arbitrary subspace. Denote by  $v_1 \in U, v_2 \in U^\perp$  the (uniquely determined) components of the vector  $v$ , i.e.  $v = v_1 + v_2$ . Then the angle  $\varphi$  between the subspace generated by  $v$  and the subspace  $U$  satisfies

$$\cos \varphi(\langle v \rangle, U) = \cos \varphi(\langle v \rangle, \langle v_1 \rangle) = \frac{\|v_1\|}{\|v\|}.$$

**PROOF.** By the Cauchy inequality,

$$\begin{aligned} \frac{|u \cdot v|}{\|u\| \|v\|} &= \frac{|u \cdot (v_1 + v_2)|}{\|u\| \|v\|} = \frac{|u \cdot v_1|}{\|u\| \|v\|} \\ &\leq \frac{\|u\| \|v_1\|}{\|u\| \|v\|} = \frac{\|v_1\|}{\|v\|} = \frac{\|v_1\|^2}{\|v\| \|v_1\|} = \frac{|v_1 \cdot v|}{\|v\| \|v_1\|}. \end{aligned}$$

**4.B.14.** A cube  $ABCD A' B' C' D'$  is given in standard notation. That is,  $ABCD$  and  $A' B' C' D'$  are faces and  $AA', BB'$  are edges. Compute the angle  $\varphi$  between  $AB'$  and  $AD'$ .

**Solution.** It can be assumed that the cube is of side 1 and placed in  $\mathbb{R}^3$  in such a way that the vertices  $A, B, C, D$  have coordinates respectively  $[0, 0, 0], [1, 0, 0], [1, 1, 0], [0, 1, 0]$  and the vertices  $A', B', C', D'$  have coordinates respectively  $[0, 0, 1], [1, 0, 1], [1, 1, 1], [0, 1, 1]$ . Thus  $AB' = B' - A = (1, 0, 1), AD' = D' - A = (0, 1, 1)$ . So

$$\cos(\varphi) = \frac{(1, 0, 1) \cdot (0, 1, 1)}{\|(1, 0, 1)\| \|(0, 1, 1)\|} = \frac{1}{2},$$

hence  $\varphi = 60^\circ$ . □

For further exercises on angles, see .

**4.B.15.** Prove that for every  $n \in \mathbb{N}$  and for all positive  $x_1, x_2, \dots, x_n \in \mathbb{R}$

$$n^2 \leq \left( \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right) \cdot (x_1 + x_2 + \dots + x_n).$$

For what arguments does equality hold?

**Solution.** It is sufficient to consider the Cauchy inequality

$$|u \cdot v| \leq \|u\| \|v\|$$

in Euclidean space  $\mathbb{R}^n$  for the vectors

$$u = \left( \frac{1}{\sqrt{x_1}}, \dots, \frac{1}{\sqrt{x_n}} \right), v = (\sqrt{x_1}, \dots, \sqrt{x_n}).$$

We get

$$(1) \quad n \leq \sqrt{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \cdot \sqrt{x_1 + x_2 + \dots + x_n}.$$

We obtain the desired inequality squaring (1). The Cauchy inequality attains equality when vector  $u$  is a multiple of  $v$ , that is, when  $x_1 = x_2 = \dots = x_n$ . □

**4.B.16.** Vectors  $\underline{u} = (u_1, u_2, u_3)$  and  $\underline{v} = (v_1, v_2, v_3)$  are given. Find a third unit vector such that parallelepiped defined by these three vectors has the greatest possible volume.

**Solution.** Denote the desired vector by  $\underline{t} = (t_1, t_2, t_3)$ . By Proposition ?? the volume of the parallelepiped  $\mathcal{P}_3(0; \underline{u}, \underline{v}, \underline{t})$  is the absolute value of determinant

$$\begin{vmatrix} u_1 & v_1 & t_1 \\ u_2 & v_2 & t_2 \\ u_3 & v_3 & t_3 \end{vmatrix} = \begin{vmatrix} t_1 & t_2 & t_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \underline{t} \cdot (\underline{u} \times \underline{v}) \leq \|\underline{t}\| \|\underline{u} \times \underline{v}\| \\ = \|\underline{u} \times \underline{v}\|.$$

The sign of the inequality follows from the Cauchy inequality. This becomes equality if and only if  $\underline{t} = c(\underline{u} \times \underline{v})$ ,  $c \in \mathbb{R}$ . The volume therefore could be at most equal to the area of paralleloid defined by vectors  $\underline{u}, \underline{v}$  (i.e. size of vector  $(\underline{u} \times \underline{v})$ ). Equality holds if and only if

$$t = \pm \frac{(\underline{u} \times \underline{v})}{\|(\underline{u} \times \underline{v})\|}.$$

for all vectors  $u \in U$ . This implies that

$$\cos \varphi(\langle v \rangle, \langle u \rangle) \leq \cos \varphi(\langle v \rangle, \langle v_1 \rangle) = \frac{\|v_1\|}{\|v\|}.$$

Thus the computed vector  $v_1$  represents the largest possible value of the cosine of angles between all choices of vectors in  $U$ . The  $\cos$  function is decreasing on the interval  $[0, \frac{\pi}{2}]$ . Hence the smallest possible angle is obtained in this way, and so the claim is proved. □

**4.1.18. Calculating angles.** The procedure in the previous lemma can be understood as follows. Choose the orthogonal projection of the one-dimensional subspace generated by  $v$  into the subspace  $U$ , and consider the ratio between  $v$  and its image. A similar procedure is used in the higher dimension. The problem is to recognize the directions whose projections give the desired (minimal) angle. This is clear in the previous example if we project the larger space  $U$  into the one-dimensional  $\langle v \rangle$  first, and then orthogonally back to  $U$ . The desired angle corresponds to the direction of the eigenvector of this map. The eigenvalue is the square of the cosine of the angle.

Let  $U_1, U_2$  be two arbitrary subspaces in a Euclidean vector space  $V$ ,  $U_1 \cap U_2 = \{0\}$ . Choose orthonormal bases  $\underline{e}$  and  $\underline{e}'$  of the whole space  $V$  such that  $U_1 = \langle e_1, \dots, e_k \rangle$ ,  $U_2 = \langle e'_1, \dots, e'_l \rangle$ .

Consider the orthogonal projection  $\varphi$  of the space  $V$  on  $U_2$ . Its restriction on  $U_1$  will be denoted by  $\varphi : U_1 \rightarrow U_2$  as before. Similarly, let  $\psi : U_2 \rightarrow U_1$  be the map which has arisen from the orthogonal projection on  $U_1$ . In the bases  $(e_1, \dots, e_k)$  and  $(e'_1, \dots, e'_l)$ , these maps have matrices  $A$  and  $B$ ,

$$\begin{pmatrix} e_1 \cdot e'_1 & \dots & e_k \cdot e'_1 \\ \vdots & & \vdots \\ e_1 \cdot e'_l & \dots & e_k \cdot e'_l \end{pmatrix}, \begin{pmatrix} e'_1 \cdot e_1 & \dots & e'_l \cdot e_1 \\ \vdots & & \vdots \\ e'_1 \cdot e_k & \dots & e'_l \cdot e_k \end{pmatrix},$$

respectively. Since  $e_i \cdot e'_j = e'_j \cdot e_i$  holds for all indices  $i, j$ ,  $B = A^T$ .

The composition of maps  $\psi \circ \varphi : U_1 \rightarrow U_1$  has therefore a symmetric positive semidefinite matrix  $A^T A$ , and  $\psi$  is adjoint to  $\varphi$ . Each such map has only nonnegative real eigenvalues. It has a diagonal matrix with these eigenvalues on the diagonal in a suitable orthonormal basis, see 3.4.7 a 3.4.9.

Now we can derive a general procedure for computing the angle  $\alpha = \varphi(U_1, U_2)$ .

**Theorem.** In the previous notation, let  $\lambda$  be the largest eigenvalue of the matrix  $A^T A$ . Then  $(\cos \alpha)^2 = \lambda$ .

**PROOF.** Let  $u \in U_1$  be the eigenvector of the map  $\psi \circ \varphi$  corresponding to the eigenvalue  $\lambda$ . Consider all eigenvalues  $\lambda_1, \dots, \lambda_k$  (including multiplicities), and let  $\underline{u} = (u_1, \dots, u_n)$  be the corresponding orthonormal basis of  $U_1$  containing the eigenvectors. Assume that  $\lambda = \lambda_1$ , and choose its eigenvector  $u = u_1$ ,  $\|u\| = 1$ . □



**4.B.17.** Find the foot of the line passing through the point  $[0, 0, 7]$  and perpendicular to the plane

$$\rho : [0, 5, 3] + (1, 2, 1)t + (-2, 1, 1)s.$$

**4.B.18.** In Euclidean space  $\mathbb{R}^5$  determine the distance between the planes

$$\begin{aligned} \varrho_1 &: [7, 2, 7, -1, 1] + t_1(1, 0, -1, 0, 0) + s_1(0, 1, 0, 0, -1), \\ \varrho_2 &: [2, 4, 7, -4, 2] + t_2(1, 1, 1, 0, 1) + s_2(0, -2, 0, 0, 3), \end{aligned}$$

where  $t_1, s_1, t_2, s_2 \in \mathbb{R}$ .

**Solution.** First compute the orthogonal complement to sum of vectors defining the planes. Form a matrix with rows as the direction vectors of the planes. Then transform this matrix into normal form.

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & -2 & 0 & 0 & 3 \end{pmatrix} \sim \dots \sim \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

So the orthogonal complement is  $\langle(0, 0, 0, 1, 0)\rangle$ . The vector  $(0, 0, 0, 1, 0)$  lies within the orthogonal complement. Transform the matrix into normal form. This shows that the orthogonal complement is one-dimensional. The distance between planes is the length of the perpendicular projection of the vector  $A_1 - A_2$  into the subspace  $\langle(0, 0, 0, 1, 0)\rangle$  for arbitrary points  $A_1 \in \varrho_1, A_2 \in \varrho_2$ . Choose e.g.  $A_1 = [7, 2, 7, -1, 1], A_2 = [2, 4, 7, -4, 2]$ . The orthogonal projection  $A_1 - A_2 = (5, -2, 0, 3, -1)$  to  $\langle(0, 0, 0, 1, 0)\rangle$  is  $(0, 0, 0, 3, 0)$ . The length of  $(0, 0, 0, 3, 0)$  gives the desired distance 3.  $\square$

**4.B.19.** In Euclidean space  $\mathbb{R}^5$  determine the distance of planes

$$\begin{aligned} \sigma_1 &: [0, 1, 2, 0, 0] + p_1(2, 1, 0, 0, 1) + q_1(-2, 0, 1, 1, 0), \\ \sigma_2 &: [3, -1, 7, 7, 3] + p_2(2, 2, 4, 0, 3) + q_2(2, 0, 0, -2, -1), \end{aligned}$$

where  $p_1, q_1, p_2, q_2 \in \mathbb{R}$ .

**Solution.** The sum of the directions  $\sigma_1, \sigma_2$  is generated by the direction vectors. Denote them by

$$\begin{aligned} u_1 &= (2, 1, 0, 0, 1), & u_2 &= (-2, 0, 1, 1, 0), \\ v_1 &= (2, 2, 4, 0, 3), & v_2 &= (2, 0, 0, -2, -1). \end{aligned}$$

Find points  $X_1 \in \sigma_1, X_2 \in \sigma_2$ , that equal the distance between  $\sigma_1$  and  $\sigma_2$ . This requires

$$\begin{aligned} X_1 - X_2 &= [0, 1, 2, 0, 0] - [3, -1, 7, 7, 3] \\ &\quad + p_1u_1 + q_1u_2 - p_2v_1 - q_2v_2 \end{aligned}$$

and

$$\begin{aligned} \langle X_1 - X_2, u_1 \rangle &= 0, & \langle X_1 - X_2, u_2 \rangle &= 0, \\ \langle X_1 - X_2, v_1 \rangle &= 0, & \langle X_1 - X_2, v_2 \rangle &= 0. \end{aligned}$$

We need to show that the angle between an arbitrary  $v \in U_1$  and  $U_2$  is at least as large as the angle between  $u$  and  $U_2$ . Equivalently that the cosine of the corresponding angle cannot be greater. By the previous lemma, it is sufficient to discuss the angle between  $u$  and  $\varphi(u) \in U_2$ . Choose  $v \in U_1, v = a_1u_1 + \dots + a_ku_k, \sum_{i=1}^k a_i^2 = \|v\|^2 = 1$ . Then

$$\begin{aligned} \|\varphi(v)\|^2 &= \varphi(v) \cdot \varphi(v) = (\psi \circ \varphi(v)) \cdot v \\ &\leq \|\psi \circ \varphi(v)\| \|v\| = \|\psi \circ \varphi(v)\|. \end{aligned}$$

Moreover, the previous lemma gives a formula for computing the angle  $\alpha$  between the vector  $v$  and the subspace  $U_2$

$$\cos \alpha = \frac{\|\varphi(v)\|}{\|v\|} = \|\varphi(v)\|.$$

Since  $\lambda_1$  is the largest eigenvalue, and the sum of squares of coordinates  $a_i^2$  is one,

$$\begin{aligned} (\cos \alpha)^2 &= \|\varphi(v)\|^2 \leq \|\psi \circ \varphi(v)\| = \sqrt{\sum_{i=1}^k (\lambda_i a_i)^2} = \\ &= \sqrt{\lambda_1^2 + \sum_{i=1}^k a_i^2 (\lambda_i^2 - \lambda_1^2)} \leq \sqrt{\lambda_1^2}. \end{aligned}$$

If  $v = u$ , we have  $\|\varphi(v)\|^2 = \lambda_1^2 \|v\|^2 = \lambda^2$ , and thus the angle has the minimal value for this vector.  $\square$

**4.1.19. Calculating volume.** An indication of how to calculate volumes in plane geometry is given at the end of the fifth part of the first chapter (see 1.5.11). There the notion of orientation played a fundamental role. We can imagine orientation as the decision whether to look at the plane  $\mathbb{R}^2$  from above or from below. The distinction lies in the order of selecting standard basis vectors  $e_1$  and  $e_2$  on the unit circle. We proceed in the same way in general:



#### ORIENTATION OF A VECTOR SPACE

Two bases  $\underline{u}$  and  $\underline{v}$  of a real vector space  $V$  are said to determine the same *orientation* if the transformation matrix between them has a positive determinant. By the orientation of a vector space  $V$  is meant the equivalence class of bases  $\underline{u}$  with respect to the equivalence defined above, by the sign of the determinant. Equivalent bases in this sense are called compatible with the chosen orientation.

It follows that there exist exactly two orientations on every vector space. From each compatible basis there is a non compatible one by a transformation matrix with a negative determinant.

A vector space with a chosen orientation is called the *oriented vector space*.

The *oriented Euclidean (point) space* is a Euclidean point space whose difference space is oriented. In the sequel we consider the standard Euclidean space  $\mathcal{E}_n$  together with the orientation given by the standard basis of  $\mathbb{R}^n$ .

Hence

$$\begin{aligned} \langle (-3, 2, -5, -7, -3), u_1 \rangle + p_1 \langle u_1, u_1 \rangle + q_1 \langle u_2, u_1 \rangle \\ - p_2 \langle v_1, u_1 \rangle - q_2 \langle v_2, u_1 \rangle &= 0, \\ \langle (-3, 2, -5, -7, -3), u_2 \rangle + p_1 \langle u_1, u_2 \rangle + q_1 \langle u_2, u_2 \rangle \\ - p_2 \langle v_1, u_2 \rangle - q_2 \langle v_2, u_2 \rangle &= 0, \\ \langle (-3, 2, -5, -7, -3), v_1 \rangle + p_1 \langle u_1, v_1 \rangle + q_1 \langle u_2, v_1 \rangle \\ - p_2 \langle v_1, v_1 \rangle - q_2 \langle v_2, v_1 \rangle &= 0, \\ \langle (-3, 2, -5, -7, -3), v_2 \rangle + p_1 \langle u_1, v_2 \rangle + q_1 \langle u_2, v_2 \rangle \\ - p_2 \langle v_1, v_2 \rangle - q_2 \langle v_2, v_2 \rangle &= 0. \end{aligned}$$

By computing the dot products, we obtain the linear equation system

$$\begin{array}{cccccc} 6p_1 & - & 4q_1 & - & 9p_2 & - & 3q_2 & = & 7, \\ -4p_1 & + & 6q_1 & & & + & 6q_2 & = & 6, \\ 9p_1 & & & - & 33p_2 & - & q_2 & = & 31, \\ 3p_1 & - & 6q_1 & - & p_2 & - & 9q_2 & = & -11. \end{array}$$

Solve it by forming a matrix and performing elementary row operations.

$$\left( \begin{array}{cccc|c} 6 & -4 & -9 & -3 & 7 \\ -4 & 6 & 0 & 6 & 6 \\ 9 & 0 & -33 & -1 & 31 \\ 3 & -6 & -1 & -9 & -11 \end{array} \right) \sim \dots \sim \left( \begin{array}{cccc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 2 \end{array} \right).$$

The solution is  $(p_1, q_1, p_2, q_2) = (0, -1, -1, 2)$ . Consequently

$$\begin{aligned} X_1 - X_2 &= (-3, 2, -5, -7, -3) - u_2 + v_1 - 2v_2 \\ &= (-3, 4, -2, -4, 2). \end{aligned}$$

The length of the vector  $(-3, 4, -2, -4, 2)$  equals the distance between the planes  $\sigma_1, \sigma_2$  and is then

$$7 = \sqrt{(-3)^2 + 4^2 + (-2)^2 + (-4)^2 + 2^2}.$$

We solved this problem by a method different to that of the previous problem. We can use both methods in both cases. Try the former method for the case of  $\sigma_1, \sigma_2$ . Find the orthogonal complement of vector subspace generated by  $(2, 1, 0, 0, 1), (-2, 0, 1, 1, 0), (2, 2, 4, 0, 3), (2, 0, 0, -2, -1)$ . We get

$$\left( \begin{array}{ccccc} 2 & 1 & 0 & 0 & 1 \\ -2 & 0 & 1 & 1 & 0 \\ 2 & 2 & 4 & 0 & 3 \\ 2 & 0 & 0 & -2 & -1 \end{array} \right) \sim \dots \sim \left( \begin{array}{ccccc} 1 & 0 & 0 & 0 & 3/2 \\ 0 & 1 & 0 & 0 & -2 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{array} \right),$$

The orthogonal complement is  $\langle (-3/2, 2, -1, -2, 1) \rangle$ , or rather  $\langle (3, -4, 2, 4, -2) \rangle$ . Note that the distance between  $\sigma_1$

Let  $u_1, \dots, u_k$  be arbitrary vectors in the difference space  $\mathbb{R}^n$ ,  $A \in \mathcal{E}_n$  a point. As an example of a convex set, the parallelepiped  $\mathcal{P}_k(A; u_1, \dots, u_k) \subset \mathcal{E}_n$  is given by

$$\mathcal{P}_k(A; u_1, \dots, u_k) = \{A + c_1 u_1 + \dots + c_k u_k; 0 \leq c_i \leq 1\}.$$

If  $u_1, \dots, u_k$  are linearly dependent, the parallelepiped is degenerated and we see the volume

$$\text{Vol } \mathcal{P}_k = 0.$$

If the vectors  $u_1, \dots, u_k$  are linearly independent, we have a  $k$ -dimensional parallelepiped  $\mathcal{P}_k(A; u_1, \dots, u_k) \subset \mathcal{E}_n$ . For given vectors  $u_1, \dots, u_k$  there are also parallelepipeds of lower dimension

$$\mathcal{P}_1(A; u_1), \dots, \mathcal{P}_k(A; u_1, \dots, u_k)$$

in Euclidean subspaces  $A + \langle u_1 \rangle, \dots, A + \langle u_1, \dots, u_k \rangle$  at our disposal. We proceed as in the case of the Gram–Schmidt orthogonalization: Consider the decomposition

$$\begin{aligned} \langle u_1, \dots, u_k \rangle &= \\ &= \langle u_1, \dots, u_{k-1} \rangle \oplus \langle u_1, \dots, u_{k-1} \rangle^\perp \cap \langle u_1, \dots, u_k \rangle. \end{aligned}$$

In this direct sum,  $u_k$  is uniquely expressed as

$$u_k = u'_k + e_k$$

where  $e_k \perp \langle u_1, \dots, u_{k-1} \rangle$ .

The absolute value of the volume of a parallelepiped is defined inductively such that it is the product of the volume of the “base” and the “altitude”:  $|\text{Vol } \mathcal{P}_1(A; u_1)| = \|u_1\|$ ,

$$\begin{aligned} |\text{Vol } \mathcal{P}_k(A; u_1, \dots, u_k)| &= \\ &= \|e_k\| |\text{Vol } \mathcal{P}_{k-1}(A; u_1, \dots, u_{k-1})|. \end{aligned}$$

If  $u_1, \dots, u_n$  is a basis compatible with the orientation of the entire vector space  $V$ , the (oriented) volume of the parallelepiped is defined by

$$\text{Vol } \mathcal{P}_n(A; u_1, \dots, u_n) = |\text{Vol } \mathcal{P}_n(A; u_1, \dots, u_n)|.$$

In the case of a non compatible basis we set

$$\text{Vol } \mathcal{P}_n(A; u_1, \dots, u_n) = -|\text{Vol } \mathcal{P}_n(A; u_1, \dots, u_n)|.$$

**Theorem.** Let  $Q \subset \mathcal{E}_n$  be a Euclidean subspace, and let  $e = (e_1, \dots, e_k)$  be an orthonormal basis of  $Z(Q)$ . For arbitrary vectors  $u_1, \dots, u_k \in Z(Q)$  and  $A \in Q$  the following holds

$$\begin{aligned} (1) \text{ Vol } \mathcal{P}_k(A; u_1, \dots, u_k) &= \begin{vmatrix} u_1 \cdot e_1 & \dots & u_k \cdot e_1 \\ \vdots & & \vdots \\ u_1 \cdot e_k & \dots & u_k \cdot e_k \end{vmatrix} \\ (2) (\text{Vol } \mathcal{P}_k(A; u_1, \dots, u_k))^2 &= \begin{vmatrix} u_1 \cdot u_1 & \dots & u_k \cdot u_1 \\ \vdots & & \vdots \\ u_1 \cdot u_k & \dots & u_k \cdot u_k \end{vmatrix} \end{aligned}$$

PROOF. The matrix

$$A = \begin{pmatrix} u_1 \cdot e_1 & \dots & u_k \cdot e_1 \\ \vdots & & \vdots \\ u_1 \cdot e_k & \dots & u_k \cdot e_k \end{pmatrix}$$

and  $\sigma_2$  equals the size of the orthogonal projection of the vector (the difference of an arbitrary point in  $\sigma_1$  and an arbitrary point in  $\sigma_2$ )

$$u = (3, -2, 5, 7, 3) = [3, -1, 7, 7, 3] - [0, 1, 2, 0, 0]$$

to this orthogonal complement. Denote the orthogonal projection of  $u$  as  $p_u$  and choose  $v = (3, -4, 2, 4, -2)$ . Obviously  $p_u = a \cdot v$  for some  $a \in \mathbb{R}$  and

$$\langle u - p_u, v \rangle = 0, \quad \text{tj.} \quad \langle u, v \rangle - a \langle v, v \rangle = 0.$$

Computing gives  $49 - a \cdot 49 = 0$ . Therefore  $p_u = 1 \cdot v = v$  and the distance between the planes  $\sigma_1$  and  $\sigma_2$  equals

$$\|p_u\| = \sqrt{3^2 + (-4)^2 + 2^2 + 4^2 + (-2)^2} = 7.$$

The method of computing the distance using the orthogonal complement of sum of vector spaces proves to be a "faster way to the solution". It is the same for the planes  $\varrho_1$  and  $\varrho_2$ . The second method however reveals points where the distance can be measured (a pair of points in which the planes are the closest). Find such points in the case of planes  $\varrho_1, \varrho_2$ . Denote

$$u_1 = (1, 0, -1, 0, 0), \quad u_2 = (0, 1, 0, 0, -1),$$

$$v_1 = (1, 1, 1, 0, 1), \quad v_2 = (0, -2, 0, 0, 3).$$

Points  $X_1 \in \varrho_1, X_2 \in \varrho_2$ , which are the "closest" (as commented above), are

$$X_1 = [7, 2, 7, -1, 1] + t_1 u_1 + s_1 u_2,$$

$$X_2 = [2, 4, 7, -4, 2] + t_2 v_1 + s_2 v_2,$$

so

$$X_1 - X_2 = [7, 2, 7, -1, 1] - [2, 4, 7, -4, 2]$$

$$+ t_1 u_1 + s_1 u_2 - t_2 v_1 - s_2 v_2$$

$$= (5, -2, 0, 3, -1)$$

$$+ t_1 u_1 + s_1 u_2 - t_2 v_1 - s_2 v_2.$$

The dot products

$$\langle X_1 - X_2, u_1 \rangle = 0, \quad \langle X_1 - X_2, u_2 \rangle = 0,$$

$$\langle X_1 - X_2, v_1 \rangle = 0, \quad \langle X_1 - X_2, v_2 \rangle = 0$$

then lead to the linear equation system

$$2t_1 = -5,$$

$$2s_1 + 5s_2 = 1,$$

$$-4t_2 - s_2 = -2,$$

$$-5s_1 - t_2 - 13s_2 = -1$$

with the unique solution  $t_1 = -5/2, s_1 = 41/2, t_2 = 5/2, s_2 = -8$ . We obtained

$$X_1 = [7, 2, 7, -1, 1] - \frac{5}{2}u_1 + \frac{41}{2}u_2$$

$$= \left[ \frac{9}{2}, \frac{45}{2}, \frac{19}{2}, -1, -\frac{39}{2} \right],$$

$$X_2 = [2, 4, 7, -4, 2] + \frac{5}{2}v_1 - 8v_2$$

$$= \left[ \frac{9}{2}, \frac{45}{2}, \frac{19}{2}, -4, -\frac{39}{2} \right].$$

The distance between the points  $X_1, X_2$  equals the distance between the planes  $\varrho_1, \varrho_2$  both of which are given by  $\|X_1 - X_2\| = \|(0, 0, 0, 3, 0)\| = 3$ .  $\square$

has the coordinates of the vectors  $u_1, \dots, u_k$  in the chosen basis in columns, and

$$|A|^2 = |A||A| = |A^T||A| = |A^T A|$$

$$= \begin{vmatrix} u_1 \cdot u_1 & \dots & u_k \cdot u_1 \\ \vdots & & \vdots \\ u_1 \cdot u_k & \dots & u_k \cdot u_k \end{vmatrix}.$$

Hence if (1) holds, then also (2) holds.

Directly from the definition, the unoriented volume equals the product

$$|\text{Vol } \mathcal{P}_k(A; u_1, \dots, u_k)| = \|v_1\| \|v_2\| \dots \|v_k\|,$$

where  $v_1 = u_1, v_2 = u_2 + a_1^2 v_1, \dots, v_k = u_k + a_1^k v_1 + \dots + a_{k-1}^k v_{k-1}$  is the result of the Gram-Schmidt orthogonalization. Thus

$$\begin{aligned} (\text{Vol } \mathcal{P}_k(A; u_1, \dots, u_k))^2 &= \begin{vmatrix} v_1 \cdot v_1 & 0 & \dots & 0 \\ \vdots & \ddots & & \\ 0 & 0 & \dots & v_k \cdot v_k \end{vmatrix} \\ &= \begin{vmatrix} v_1 \cdot v_1 & \dots & v_k \cdot v_1 \\ \vdots & & \vdots \\ v_1 \cdot v_k & \dots & v_k \cdot v_k \end{vmatrix}. \end{aligned}$$

Denote by  $B$  the matrix whose columns are formed by the coordinates of vectors  $v_1, \dots, v_k$  in the orthonormal basis  $\underline{e}$ . Since  $v_1, \dots, v_k$  have arisen from  $u_1, \dots, u_k$  as images under a linear transformation with an upper-triangular matrix  $C$  with ones on the diagonal,  $B = CA$  and  $|B| = |C||A| = |A|$ . But then  $|A|^2 = |B|^2 = |A||A|$ , and thus  $\text{Vol } \mathcal{P}_k(A; u_1, \dots, u_k) = \pm|A|$ . The resulting volume is zero if the vectors  $u_1, \dots, u_k$  are linearly dependent. Provided they are independent, the sign of the determinant is positive if and only if the basis  $u_1, \dots, u_k$  defines the same orientation as the basis  $\underline{e}$ .  $\square$

Consider a parallelepiped in a  $k$ -dimensional space, which is spanned by  $k$  vectors. Write down the coordinates (in an orthonormal basis) into the columns of a matrix. Then the volume of the parallelepiped is the determinant of the matrix.

The formula (2) above is called the *Gram determinant*. It is independent of the choice of basis and, therefore it is useful when  $k$  is lower than the dimension of the whole space.

We formulate the following important geometric consequence:

**4.1.20. Corollary.** *For each linear map  $\varphi : V \rightarrow V$  on a Euclidean vector space  $V$ ,  $\det \varphi$  equals the (oriented) volume of the image of the parallelepiped determined by vectors of an orthonormal basis. More generally, the image of the parallelepiped  $\mathcal{P}$ , determined by arbitrary  $\dim V$  vectors, has a volume equal to  $\det \varphi$ -multiple of the former volume.*

**4.B.20.** Find the intersection of the plane passing through the point  $A = [1, 2, 3, 4] \in \mathbb{R}^4$  and orthogonal to the plane

$$\varrho : [1, 0, 1, 0] + (1, 2, -1, -2)s + (1, 0, 0, 1)t, \quad s, t \in \mathbb{R}.$$

**Solution.** Find the plane orthogonal to  $\varrho$ . Its direction is orthogonal to the direction of  $\varrho$ , for vectors  $(a, b, c, d)$  within its direction we get linear equation system

$$\begin{aligned} (a, b, c, d) \cdot (1, 2, -1, -2) &= 0 \quad \equiv \quad a + 2b - c - 2d = 0 \\ (a, b, c, d) \cdot (1, 0, 0, 1) &= 0 \quad \equiv \quad a + d = 0. \end{aligned}$$

The solution is the two-dimensional vector space  $\langle (0, 1, 2, 0), (-1, 0, -3, 1) \rangle$ . The plane  $\tau$  orthogonal to  $\varrho$  passing through  $A$  has parametric equation

$$\tau : [1, 2, 3, 4] + (0, 1, 2, 0)u + (-1, 0, -3, 1)v, \quad u, v \in \mathbb{R}.$$

We can obtain the intersection of the planes from both parametric equations. It is given by the linear equation system

$$\begin{aligned} 1 + s + t &= 1 - v \\ 2s &= 2 + u \\ 1 - s &= 3 + 2u - 3v \\ -2s + t &= 4 + v, \end{aligned}$$

which has the unique solution (it must be so as matrix columns are linearly independent)  $s = -8/19$ ,  $t = 34/19$ ,  $u = -54/19$ ,  $v = -26/19$ . Substitute the parameter values  $s$  and  $t$  into the parametric form of the plane  $\varrho$ , to obtain the intersection  $[45/19, -16/19, 11/19, 18/19]$ . (Needless to say, the same solution is obtained by substituting the values into  $\tau$ ).  $\square$

**4.B.21.** Find a line passing through point  $[1, 2] \in \mathbb{R}^2$  so that the angle between this line and the line

$$p : [0, 1] + t(1, 1)$$

is  $30^\circ$ .

**Solution.** The angle between two lines is the angle between their direction vectors. It is sufficient to find the direction vector  $\underline{v}$  of the line. One way to do so is to rotate the direction vector of  $p$  by  $30^\circ$ . The rotation matrix for the angle  $30^\circ$  is

$$\begin{pmatrix} \cos 30^\circ & -\sin 30^\circ \\ \sin 30^\circ & \cos 30^\circ \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}.$$

The desired vector  $\underline{v}$  is therefore

$$\underline{v} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} - \frac{1}{2} \\ \frac{\sqrt{3}}{2} + \frac{1}{2} \end{pmatrix}.$$

We could perform the backward rotation as well. The line (one of two possible) has parametric equation

$$[1, 2] + \left( \frac{\sqrt{3}}{2} - \frac{1}{2}, \frac{\sqrt{3}}{2} + \frac{1}{2} \right) t.$$

$\square$

**4.1.21. Outer product and cross product.** The previous considerations are closely related to the tensor product of vectors. We do not go further in this technically more complicated topic now. But we do mention the outer product of  $n = \dim V$  vectors  $u_1, \dots, u_n \in V$ .



Let  $(u_{1j}, \dots, u_{nj})^T$  be coordinate expressions of vectors  $u_j$  in a chosen orthonormal basis  $V$ . Let  $M$  be a matrix with elements  $(u_{ij})$ . Then the determinant  $|M|$  does not depend on the choice of the basis with the same orientation. Its value is called the *outer product* of the vectors  $u_1, \dots, u_n$ , and is denoted by  $[u_1, \dots, u_n]$ . Hence the outer product is the oriented volume of the corresponding parallelepiped, see 4.1.19.

Although the outer product looks like a scalar quantity, the story gets more complicated once we allow for general basis of  $V$ . Then the determinant of the matrix  $M$  built of the coordinates of  $u_i$  changes by the determinants of the transition matrices. Such objects are called *densities* and we shall come back to them in chapter 9.

Several useful properties of the outer product follow directly from the definition

- (1) The map  $(u_1, \dots, u_n) \mapsto [u_1, \dots, u_n]$  is an antisymmetric  $n$ -linear map. It is linear in all arguments, and the interchange of any two arguments causes a change of sign.
- (2) The outer product is zero if and only if the vectors  $u_1, \dots, u_n$  are linearly dependent.
- (3) The vectors  $u_1, \dots, u_n$  form a positive basis if and only if the outer product is positive.

Consider a Euclidean vector space  $V$  of dimension  $n \geq 2$  and vectors  $u_1, \dots, u_{n-1} \in V$ . If these  $n-1$  vectors are substituted into the first  $n-1$  arguments of the  $n$ -linear map defined by the volume determinant as above, then there is one argument left over. This defines a linear form on  $V$ . Since the scalar product is available, each linear form corresponds to exactly one vector. This vector  $v \in V$  is called the *cross product* of the vectors  $u_1, \dots, u_{n-1}$ . For each vector  $w \in V$

$$\langle w, v \rangle = [u_1, \dots, u_{n-1}, w].$$

We denote the cross product by  $v = u_1 \times \dots \times u_{n-1}$ .

If the coordinates of the vectors in an orthonormal basis are  $v = (y_1, \dots, y_n)^T$ ,  $w = (x_1, \dots, x_n)^T$  and  $u_j = (u_{1j}, \dots, u_{nj})^T$ , then the definition can be expressed as

$$y_1 x_1 + \dots + y_n x_n = \begin{vmatrix} u_{11} & \dots & u_{1(n-1)} & x_1 \\ \vdots & & \vdots & \vdots \\ u_{n1} & \dots & u_{n(n-1)} & x_n \end{vmatrix}.$$

Hence the vector  $v$  is determined uniquely. Its coordinates are calculated by the formal expansion of this determinant along the last column. The following properties of the cross product are direct consequences of the definition:

**Theorem.** For the cross product  $v = u_1 \times \dots \times u_{n-1}$

- (1)  $v \in \langle u_1, \dots, u_{n-1} \rangle^\perp$
- (2)  $v$  is nonzero if and only if the vectors  $u_1, \dots, u_{n-1}$  are linearly independent,

**4.B.22.** An octahedron has eight faces consisting of equilateral triangles. Determine  $\cos \alpha$ , where  $\alpha$  is the angle between two adjacent faces of a regular octahedron.

**Solution.** An octahedron is symmetric, therefore it does not matter which two faces are selected. By suitable scaling, the octahedron has edge length 1 and is placed in the standard Cartesian coordinate system  $\mathbb{R}^3$  so that its centroid is at  $[0, 0, 0]$ . Its vertices then are located at the points  $A = [\frac{\sqrt{2}}{2}, 0, 0]$ ,  $B = [0, \frac{\sqrt{2}}{2}, 0]$ ,  $C = [-\frac{\sqrt{2}}{2}, 0, 0]$ ,  $D = [0, -\frac{\sqrt{2}}{2}, 0]$ ,  $E = [0, 0, -\frac{\sqrt{2}}{2}]$  and  $F = [0, 0, \frac{\sqrt{2}}{2}]$ .

We compute the angle between the faces  $CDF$  and  $BCF$ . We need to find vectors orthogonal to their intersection and lying within respective faces, which means orthogonal to  $CF$ . They are altitudes from  $D$  and  $F$  to the edge  $CF$  in the triangles  $CDF$  and  $BCF$  respectively. The altitudes in an equilateral triangle are the same segments as the medians, so they are  $SD$  and  $SB$ , where  $S$  is midpoint of  $CF$ . Because the coordinates of points  $C$  and  $F$  are known,  $S$  has coordinates  $[-\frac{\sqrt{2}}{4}, 0, \frac{\sqrt{2}}{4}]$  and the vectors are  $SD = (\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})$  and  $SB = (\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})$ . Together

$$\cos \alpha = \frac{(\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4}) \cdot (\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})}{\|(\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})\| \|(\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{4})\|}} = -\frac{1}{3}.$$

Therefore  $\alpha \doteq 132^\circ$ . □

**4.B.23.** In Euclidean space  $\mathbb{R}^5$  determine the angle  $\varphi$  between subspaces  $U, V$ , where

- (a)  $U : [3, 5, 1, 7, 2] + t(1, 0, 2, -2, 1), t \in \mathbb{R}$ ,  
 $V : [0, 1, 0, 0, 0] + s(2, 0, -2, 1, -1), s \in \mathbb{R}$ ;
- (b)  $U : [4, 1, 1, 0, 1] + t(2, 0, 0, 2, 1), t \in \mathbb{R}$ ,  
 $V : x_1 + x_2 + x_3 + x_5 = 7$ ;
- (c)  $U : 2x_1 - x_2 + 2x_3 + x_5 = 3$ ,  
 $V : x_1 + 2x_2 + 2x_3 + x_5 = -1$ ;
- (d)  $U : [0, 1, 1, 0, 0] + t(0, 0, 0, 1, -1), t \in \mathbb{R}$ ,  
 $V : [1, 0, 1, 1, 1] + r(1, -1, 2, 1, 0) + s(0, 1, 3, 2, 0)$   
 $+ p(1, 0, 0, 1, 0) + q(1, 3, 1, 0, 0)$ ,  
 $r, s, p, q \in \mathbb{R}$ ;
- (e)  $U : [0, 2, 5, 0, 0] + t(2, 1, 3, 5, 3) + s(0, 3, 1, 4, -2)$   
 $+ r(1, 2, 4, 0, 3), t, s, r \in \mathbb{R}$ ,  
 $V : [0, 0, 0, 0, 0] + p(-1, 1, 1, -5, 0)$   
 $+ q(1, 5, 1, 13, -4), p, q \in \mathbb{R}$ ;
- (f)  $U : [1, 1, 1, 1, 1] + t(1, 0, 1, 1, 1)$   
 $+ s(1, 0, 0, 1, 1), t, s \in \mathbb{R}$ ,  
 $V : [1, 1, 1, 1, 1] + p(1, 1, 1, 1, 1) + q(1, 1, 0, 1, 1)$   
 $+ r(1, 1, 0, 1, 0), p, q, r \in \mathbb{R}$ .

**Solution.** Recall that the angle between affine subspaces is the same as the angle between vector spaces associated to them. Therefore the translation caused by the point addition can be omitted.

Case (a). Since  $U$  and  $V$  are one-dimensional spaces, the angle  $\varphi \in [0, \pi/2]$  is given by formula

$$\cos \varphi = \frac{|(1, 0, 2, -2, 1) \cdot (2, 0, -2, 1, -1)|}{\|(1, 0, 2, -2, 1)\| \|(2, 0, -2, 1, -1)\|} = \frac{5}{\sqrt{10} \cdot \sqrt{10}}.$$

- (3) the length  $\|v\|$  of the cross product equals the absolute value of the volume of parallelepiped  $\mathcal{P}(0; u_1, \dots, u_{n-1})$ ,
- (4)  $(u_1, \dots, u_{n-1}, v)$  is a compatible basis of the oriented Euclidean space  $V$ .



**PROOF.** The first claim follows directly from the defining formula for  $v$ . Substituting an arbitrary vector  $u_j$  for  $w$  gives the scalar product  $v \cdot u_j$  on the left and the determinant with two equal columns on the right.

The rank of the matrix with  $n - 1$  columns  $u_j$  is given by the maximal size of a non-zero minor. The minors which define coordinates of the cross product are of degree  $n - 1$  and thus claim (2) is proved.

If the vectors  $u_1, \dots, u_{n-1}$  are linearly dependent, then (3) also holds. Suppose the vectors are linearly independent. Let  $v$  be their cross product, and choose an orthonormal basis  $(e_1, \dots, e_{n-1})$  of the space  $\langle u_1, \dots, u_{n-1} \rangle$ . It follows from what is proved that there exists a multiple  $(1/\alpha)v, 0 \neq \alpha \in \mathbb{R}$ , such that  $(e_1, \dots, e_k, (1/\alpha)v)$  is an orthonormal basis of  $V$ . The coordinates of the vectors in this basis are

$$u_j = (u_{1j}, \dots, u_{(n-1)j}, 0)^T, \quad v = (0, \dots, 0, \alpha)^T.$$

So the outer product  $[u_1, \dots, u_{n-1}, v]$  equals (see the definition of cross product)

$$[u_1, \dots, u_{n-1}, v] = \begin{vmatrix} u_{11} & \dots & u_{1(n-1)} & 0 \\ \vdots & & \vdots & \vdots \\ u_{(n-1)1} & \dots & u_{(n-1)(n-1)} & 0 \\ 0 & \dots & 0 & \alpha \end{vmatrix} = \langle v, v \rangle = \alpha^2.$$

By expanding the determinant along the last column,

$$\alpha^2 = \alpha \text{Vol } \mathcal{P}(0; u_1, \dots, u_{n-1}),$$

which proves the remaining two claims. □

In technical applications in  $\mathbb{R}^3$ , the cross product is often used. It assigns a vector to any pair of vectors.

## 2. Transformations

This short section backs up a quite wide area of practical considerations displayed in the other column. As usual, we can understand objects well only if we master the mappings preserving the crucial concepts.



Therefore  $\cos \varphi = 1/2$  and  $\varphi = \pi/3$ .

Case (b). The subspace  $U$  has direction vector  $(2, 0, 0, 2, 1)$  and the subspace  $V$  has normal vector  $(1, 1, 1, 0, 1)$ . The angle between them  $\psi = \pi/3$  is derived from the formula

$$\cos \psi = \frac{(2,0,0,2,1) \cdot (1,1,1,0,1)}{\|(2,0,0,2,1)\| \cdot \|(1,1,1,0,1)\|} = \frac{3}{3 \cdot 2}.$$

Notice that  $\varphi = \pi/2 - \psi = \pi/6$ , because  $\varphi$  is complement to  $\psi$ .

Case (c). The hyperplanes  $U$  and  $V$  are defined by normal vectors  $u = (2, -1, 2, 0, 1)$  and  $v = (1, 2, 2, 0, 1)$ . The angle  $\varphi$  equals to angle between the direction vectors  $u$  a  $v$ . Therefore (see (a))

$$\cos \varphi = \frac{|(2,-1,2,0,1) \cdot (1,2,2,0,1)|}{\|(2,-1,2,0,1)\| \cdot \|(1,2,2,0,1)\|} = \frac{1}{2}, \quad \text{tj.} \quad \varphi = \frac{\pi}{3}.$$

Case (d). Denote

$u = (0, 0, 0, 1, -1)$ ,  $v_1 = (1, -1, 2, 1, 0)$ ,  
 $v_2 = (0, 1, 3, 2, 0)$ ,  $v_3 = (1, 0, 0, 1, 0)$ ,  $v_4 = (1, 3, 1, 0, 0)$   
 and denote the orthogonal projection of  $u$  into the vector subspace of  $V$  (subspace generated by  $v_1, v_2, v_3, v_4$ ) by  $p_u$ .

Now

$$p_u = av_1 + bv_2 + cv_3 + dv_4 \quad \text{for some } a, b, c, d \in \mathbb{R}$$

and

$$\begin{aligned} \langle p_u - u, v_1 \rangle &= 0, & \langle p_u - u, v_2 \rangle &= 0, \\ \langle p_u - u, v_3 \rangle &= 0, & \langle p_u - u, v_4 \rangle &= 0. \end{aligned}$$

Substituting for  $p_u$  gives the linear equation system

$$\begin{aligned} 7a + 7b + 2c &= 1, \\ 7a + 14b + 2c + 6d &= 2, \\ 2a + 2b + 2c + d &= 1, \\ 6b + c + 11d &= 0. \end{aligned}$$

The solution is  $(a, b, c, d) = (-8/19, 7/19, 13/19, -5/19)$ .

$$(1) \quad \cos \varphi = \frac{\|p_u\|}{\|u\|}$$

and so

$$\begin{aligned} p_u &= -\frac{8}{19}v_1 + \frac{7}{19}v_2 + \frac{13}{19}v_3 - \frac{5}{19}v_4 = (0, 0, 0, 1, 0), \\ \cos \varphi &= \frac{\|p_u\|}{\|u\|} = \frac{\|(0, 0, 0, 1, 0)\|}{\|(0, 0, 0, 1, -1)\|} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}. \end{aligned}$$

Hence  $\varphi = \pi/4$ .

Case (e). Determine the intersection of the vector subspaces associated with the given affine subspaces. The vector  $(x_1, x_2, x_3, x_4, x_5)$  is in the vector subspace of  $U$ , if and only if

$(x_1, x_2, x_3, x_4, x_5) = t(2, 1, 3, 5, 3) + s(0, 3, 1, 4, -2) + r(1, 2, 4, 0, 3)$  for some  $t, s, r \in \mathbb{R}$ . Similarly,  $(x_1, x_2, x_3, x_4, x_5) \in V$  if and only if  $(x_1, x_2, x_3, x_4, x_5) = p(-1, 1, 1, -5, 0) + q(1, 5, 1, 13, -4)$  for some  $p, q \in \mathbb{R}$ . We look for such  $t, s, r, p, q \in \mathbb{R}$ , so that

$$\begin{aligned} t(2, 1, 3, 5, 3) + s(0, 3, 1, 4, -2) + r(1, 2, 4, 0, 3) \\ = p(-1, 1, 1, -5, 0) + q(1, 5, 1, 13, -4). \end{aligned}$$

It is a homogeneous linear equation system. It is solved in matrix form (order of variables is  $t, s, r, p, q$ )

AFFINE MAPS

**4.2.1.** A map  $f : \mathcal{A} \rightarrow \mathcal{B}$  between affine spaces is called an *affine map* if there exists a linear map  $\varphi : Z(\mathcal{A}) \rightarrow Z(\mathcal{B})$  between their difference spaces such that for all  $A \in \mathcal{A}$ ,  $v \in Z(\mathcal{A})$  the following holds:

$$f(A + v) = f(A) + \varphi(v).$$

The maps  $f$  and  $\varphi$  are determined uniquely by this property, and by arbitrarily chosen images of  $(\dim \mathcal{A} + 1)$  points in general position.

For an arbitrary affine combination of points  $t_0A_0 + \dots + t_sA_s \in \mathcal{A}$  we obtain

$$\begin{aligned} f(t_0A_0 + \dots + t_sA_s) &= \\ &= f(A_0 + t_1(A_1 - A_0) + \dots + t_s(A_s - A_0)) \\ &= f(A_0) + t_1\varphi(A_1 - A_0) + \dots + t_s\varphi(A_s - A_0) \\ &= t_0f(A_0) + t_1f(A_1) + \dots + t_sf(A_s). \end{aligned}$$

On the other hand, if a map preserves affine combinations, we may use a specific combination of  $n + 1$  fixed vectors generating the affine frame. After choosing successively the only non-zero coefficients  $t_i = 1$ ,  $i = 1, \dots, s$ , we define the map  $\varphi$  between difference spaces by the relation  $\varphi(A_i - A_0) = f(A_i) - f(A_0)$ . The previous computation can be read in the opposite direction, so we can check the validity and linearity of  $\varphi$ . The assumption that the first and the last rows are equal implies that the second and the third rows are equal. So we have an affine map with the corresponding linear map  $\varphi$  between difference spaces which we described in the chosen affine frame by this procedure. Therefore:

**Theorem.** *Affine maps are exactly those maps which preserve the affine combinations of points.*

It is sufficient to check the invariance of affine combinations for all pairs of points since we can create an arbitrary affine combination from them. The affine combination of  $k + 2$  points  $A_0, \dots, A_{k+1}$  can be expressed as

$$r(t_0A_0 + \dots + t_kA_k) + sA_{k+1},$$

where  $\sum_{i=0}^k t_k = 1$  and  $r + s = 1$ . We choose a point which is an affine combination of  $k + 1$  points only. Then make its combination with the last one. In this way, any finite affine combination can be made step by step from the combination of pairs.

**4.2.2. Ratio of collinear points.** The affine combinations of pairs of points can be also expressed with the help of the *ratio of points* on a straight line. If  $C$  is given by an affine combination of points  $A$  and  $B \neq C$ ,  $C = rA + sB$ , then we say that the number

$$\lambda = (C; A, B) = -\frac{s}{r}$$

is the ratio of the point  $C$  with respect to the given points  $A$  and  $B$ . Since we can express  $C$  as

$$C = A + s(B - A) = B + r(A - B),$$





$$\begin{pmatrix} 2 & 0 & 1 & 1 & -1 \\ 1 & 3 & 2 & -1 & -5 \\ 3 & 1 & 4 & -1 & -1 \\ 5 & 4 & 0 & 5 & -13 \\ 3 & -2 & 3 & 0 & 4 \end{pmatrix} \sim \dots \\ \sim \begin{pmatrix} 1 & 3 & 2 & -1 & -5 \\ 0 & 2 & 1 & -1 & -3 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The vectors defining  $V$  are linear combination of the vectors of  $U$ . So  $V$  is subset of  $U$ , and hence  $\varphi = 0$ .

Case (f). Find the intersection of  $U$  and  $V$ . Search for numbers  $t, s, p, q, r \in \mathbb{R}$  such that

$$t(1, 0, 1, 1, 1) + s(1, 0, 0, 1, 1) \\ = p(1, 1, 1, 1, 1) + q(1, 1, 0, 1, 1) + r(1, 1, 0, 1, 0).$$

The solution is  $(t, s, p, q, r) = (-a, a, -a, a, 0)$ ,  $a \in \mathbb{R}$ . The intersection  $Z(U) \cap Z(V)$  of vector spaces  $U$  and  $V$  contains the vectors

$$(0, 0, -a, 0, 0) = -a(1, 0, 1, 1, 1) + a(1, 0, 0, 1, 1) \\ = -a(1, 1, 1, 1, 1) + a(1, 1, 0, 1, 1) \\ + 0(1, 1, 0, 1, 0),$$

where  $a \in \mathbb{R}$ .  $Z(U) \cap Z(V)$  is generated by  $(0, 0, 1, 0, 0)$  and its orthogonal complement  $(Z(U) \cap Z(V))^\perp$  is generated by vectors  $(1, 0, 0, 0, 0)$ ,  $(0, 1, 0, 0, 0)$ ,  $(0, 0, 0, 1, 0)$ ,  $(0, 0, 0, 0, 1)$ .

We obtain

$$Z(U) \cap Z(V) \neq \{0\}, \quad Z(U) \cap Z(V) \neq Z(U), \\ Z(U) \cap Z(V) \neq Z(V).$$

The angle  $\varphi$  is defined as the angle between the subspaces  $Z(U) \cap (Z(U) \cap Z(V))^\perp$  and  $Z(V) \cap (Z(U) \cap Z(V))^\perp$ .

It is now established that

$$Z(U) \cap (Z(U) \cap Z(V))^\perp = \langle (1, 0, 0, 1, 1) \rangle, \\ Z(V) \cap (Z(U) \cap Z(V))^\perp = \langle (1, 1, 0, 1, 1), (1, 1, 0, 1, 0) \rangle.$$

It is enough to express  $Z(U)$  as a linear combination of vectors  $(0, 0, 1, 0, 0)$ ,  $(1, 0, 0, 1, 1)$  and  $Z(V)$  by the vectors  $(0, 0, 1, 0, 0)$ ,  $(1, 1, 0, 1, 1)$ ,  $(1, 1, 0, 1, 0)$ . Since the dimension of  $Z(U) \cap (Z(U) \cap Z(V))^\perp$  is 1, we can use the formula (1), where  $u = (1, 0, 0, 1, 1)$  and  $p_u$  is the orthogonal projection of  $u$  into  $Z(V) \cap (Z(U) \cap Z(V))^\perp$ . Then

$$p_u = a(1, 1, 0, 1, 1) + b(1, 1, 0, 1, 0)$$

and

$$\langle p_u - u, (1, 1, 0, 1, 1) \rangle = 0, \quad \langle p_u - u, (1, 1, 0, 1, 0) \rangle = 0,$$

which leads to the linear equation system

$$4a + 3b = 3, \\ 3a + 3b = 2$$

with the unique solution  $a = 1$ ,  $b = -1/3$ . Thus

$$p_u = \left(\frac{2}{3}, \frac{2}{3}, 0, \frac{2}{3}, 1\right).$$

From (1) it follows that

the ratio  $\lambda$  is the ratio of the length of the oriented vectors  $C - A$  and  $C - B$ . In particular,  $\lambda = -1$  if and only if  $C$  is at the centre of the line segment joining  $A$  and  $B$  (i.e.  $r = s = \frac{1}{2}$  in the affine combination).

Hence the characterization of affine maps in terms of affine combinations has the following consequence:

**Corollary.** *Affine maps are exactly those maps for which the ratios are invariant.*

Needless to say that collinearity of points must be preserved in order to talk about the ratios.

**4.2.3. Coordinate expression for maps.** View a general affine map  $f : \mathcal{A} \rightarrow \mathcal{B}$ ,  $f(X) = f(A_0) + \varphi(X - X_0)$ , in coordinates as follows. First express the image  $f(A_0)$  of the origin of the frame  $(A_0, \underline{u})$  on  $\mathcal{A}$  in the frame  $(B_0, \underline{v})$  on  $\mathcal{B}$ . In other words, the vector  $f(A_0) - B_0$  has got coordinates  $y_0$  in the basis  $\underline{v}$ . Everything else is then given by multiplying by the matrix of the map  $\varphi$  in the chosen bases and by adding the outcome. Each affine map therefore has the following form in coordinates:

$$x \mapsto y_0 + Y \cdot x,$$

where  $y_0$  is as above, and  $Y$  is the matrix of the map  $\varphi$ .

Of course, the changes of coordinates are special instances of invertible affine mappings on the standard affine space  $\mathcal{A}_n$ . Similarly to the case of linear mappings, under the choice of two affine coordinate systems  $(A_0, \underline{u})$  and  $(B_0, \underline{v})$  on  $\mathcal{A}$ , the coordinate expression of the identity mapping is the requested rule for the change of coordinates, cf. 4.1.5.

Next, consider the change of the frame

$$x = w + M \cdot x',$$

on the domain by a translation  $w$  and a matrix  $M$ . Further, let

$$y' = z + N \cdot y$$

describe a change of basis on the range space by a translation  $z$  and a matrix  $N$ . Then the coordinate expression  $x \mapsto y_0 + Y \cdot x$  for  $f : \mathcal{A} \rightarrow \mathcal{B}$  is

$$(1) \quad y' = z + N \cdot y = z + N \cdot (y_0 + Y \cdot x) \\ = (z + N \cdot y_0 + N \cdot Y \cdot w) + (N \cdot Y \cdot M) \cdot x'.$$

Hence the affine map in the new bases is given by the translation vector  $z + N \cdot y_0 + N \cdot Y \cdot w$  and matrix  $N \cdot Y \cdot M$ .

#### EUCLIDEAN MAPS

**4.2.4.** The Euclidean maps  $f : \mathcal{E}_1 \rightarrow \mathcal{E}_2$  are affine maps which respects the distances, which happens if and only if the associated linear maps  $\varphi$  are orthogonal.

In particular, the coordinate description of invertible Euclidean maps includes orthogonal matrices  $Y$ . If the dimension of the image is bigger, then we always can complete the image of a chosen orthogonal frame into an orthonormal frame of the co-domain, and then the relevant matrix  $Y$  will contain an orthogonal block, completed by zeros.

$$\cos \varphi = \frac{\| (2/3, 2/3, 0, 2/3, 1) \|}{\| (1, 0, 0, 1, 1) \|} = \frac{\sqrt{7}}{3}. \quad \varphi \doteq 0.49 \ (\approx 28^\circ).$$

□

### C. Geometry of quadratic forms

**4.C.1.** Determine the polar basis of the form  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x_1, x_2, x_3) = 3x_1^2 + 2x_1x_2 + x_2^2 + 4x_2x_3 + 6x_3^2$ .

**Solution.** Its matrix is

$$A = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 1 & 2 \\ 0 & 2 & 6 \end{pmatrix}.$$

According to step (1) of the Lagrange algorithm (see Theorem 4.3.1), perform the following operations

$$\begin{aligned} f(x_1, x_2, x_3) &= \frac{1}{3}(3x_1 + x_2)^2 + \frac{2}{3}x_2^2 + 4x_2x_3 + 6x_3^2 \\ &= \frac{1}{3}y_1^2 + \frac{3}{2}\left(\frac{2}{3}y_2 + 2y_3\right)^2 \\ &= \frac{1}{3}z_1^2 + \frac{3}{2}z_2^2. \end{aligned}$$

The form has rank 2 and the matrix changing the basis to the polar basis  $\underline{w}$  is obtained by a combination of following transformations:  $z_3 = y_3 = x_3$ ,  $z_2 = \frac{2}{3}y_2 + 2y_3 = \frac{2}{3}x_2 + 2x_3$  and  $z_1 = y_1 = 3x_1 + x_2$ , so the change of basis matrix is

$$T = \begin{pmatrix} 3 & 1 & 0 \\ 0 & \frac{2}{3} & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

We computed the polar coordinates, expressed them in standard basis and wrote them as rows of the matrix (the columns of this matrix are vectors of the standard basis in the polar basis). The polar basis vector coordinates are the columns of the matrix  $T^{-1}$ .

$$T^{-1} = \begin{pmatrix} \frac{1}{3} & \frac{-1}{2} & 1 \\ 0 & \frac{3}{2} & -3 \\ 0 & 0 & 1 \end{pmatrix},$$

The polar basis is therefore  $((\frac{1}{3}, 0, 0), (-\frac{1}{2}, \frac{3}{2}, 0), (1, -3, 1))$ . □

**4.C.2.** Determine the polar basis of the form  $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ ,  $f(x_1, x_2, x_3) = 2x_1x_3 + x_2^2$ .

**Solution.** The matrix is of the form

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Change the order of the variables:  $y_1 = x_2$ ,  $y_2 = x_1$ ,  $y_3 = x_3$ . It is then trivial to apply step (1) of Lagrange algorithm (there are no common terms). However for the next step, case (4) sets in. Introduce the transformation  $z_1 = y_1$ ,  $z_2 = y_2$ ,  $z_3 = y_3 - y_2$ .

$$f(x_1, x_2, x_3) = z_1^2 + 2z_2(z_3 + z_2) = z_1^2 + \frac{1}{2}(2z_2 + z_3)^2 - \frac{1}{2}z_3^2.$$

**Remark.** Notice an amazing fact about transformations on Euclidean spaces keeping the distances invariant, i.e., general mappings  $f : \mathcal{E}_n \rightarrow \mathcal{E}_n$ .



Notice that a point  $C$  is on the segment determined by points  $A$  and  $B$  if and only if  $\|C - B\| + \|C - A\| = \|B - A\|$ . Consequently, any mapping  $f$  preserving distances must map segments to segments, thus also lines to lines. Moreover, if preserving distances, it clearly preserves ratios too, and thus we deal with an affine map. Affine maps preserve distances if and only if they are Euclidean.

In 4.D.1, we can see a purely analytic proof that in principle all such (reasonably “smooth”) maps are affine and thus Euclidean. The advantage of the above synthetic approach is, that even the smoothness of the mapping follows from the assumption.

**4.2.5. Affine and Euclidean properties.** Now we can consider which properties are related to the affine structure and where we really need the scalar product on the difference vector space.



All Euclidean transformations, i.e. bijective affine maps which preserve the distance between points, preserve also all objects we have studied (possibly up to the change of orientation). In particular, they preserve unoriented angles, unoriented volumes, angles between subspaces etc. If we want them to preserve also oriented angles, cross products, and volumes, then we must also assume that the transformations preserve the orientation.

Dealing with a general affine transformation  $T$  on an  $m$ -dimensional Euclidean space, the volumes are multiplied by the determinants of the relevant linear parts of  $T$ .

We ask: *Which concepts are preserved under affine transformations?*

Recall first that an affine transformation on an  $n$ -dimensional space  $\mathcal{A}$  is uniquely defined by mapping  $n + 1$  points in general position, that is, by mapping a one  $n$ -dimensional simplex. In the plane, this means choosing the image of any nondegenerate triangle. Preserved properties are properties related to subspaces and ratios. In particular, incidence properties of the type “a line passing through a point” or “a plane contains a line” etc. are preserved. Moreover, the collinearity of vectors is preserved. For every two collinear vectors, the ratio of their lengths is preserved independently of the scalar product defining the length. Similarly, the ratio of the volumes of two  $n$ -dimensional parallelepipeds is preserved under the transformations, since the determinant of the corresponding matrix changes them by the same multiple.

These affine properties can be used to prove geometric statements. For instance in the plane, to prove the fact that the medians of a triangle intersect in a single point, and in one third of their lengths, it is sufficient to verify this only in the case of an isosceles right-angled triangle or only in the

Together,  $z_1 = y_1 = x_2$ ,  $z_2 = y_2 = x_1$ ,  $z_3 = y_3 - y_2 = x_3 - x_1$ . The matrix  $T$  for change to polar basis is

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad T^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

The polar basis is therefore  $((0, 1, 0), (1, 0, 1), (0, 1, 1))$ .  $\square$

**4.C.3.** Find the polar basis of the quadratic form  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ , which in the standard basis is defined as

$$f(x_1, x_2, x_3) = x_1x_2 + x_1x_3.$$

**Solution.** By an application of the Lagrange algorithm:

$$\begin{aligned} f(x_1, x_2, x_3) &= 2x_1x_2 + x_2x_3 \\ \text{substitution } y_2 &= x_2 - x_1, y_1 = x_1, y_3 = x_3 \\ &= 2x_1(x_1 + y_2) + (x_1 + y_2)x_3 \\ &= 2x_1^2 + 2x_1y_2 + x_1x_3 + y_2x_3 \\ &= \frac{1}{2}(2x_1 + y_2 + \frac{1}{2}x_3)^2 - \frac{1}{2}y_2^2 - \frac{1}{8}x_3^2 + y_2x_3 \\ \text{substitution } y_1 &= 2x_1 + y_2 + \frac{1}{2}x_3 \\ &= \frac{1}{2}y_1^2 - \frac{1}{2}y_2^2 - \frac{1}{8}x_3^2 + y_2x_3 \\ &= \frac{1}{2}y_1^2 - 2(\frac{1}{2}y_2 - \frac{1}{2}x_3)^2 + \frac{3}{8}x_3^2 \\ \text{substitution } y_3 &= \frac{1}{2}y_2 - \frac{1}{2}x_3 \\ &= \frac{1}{2}y_1^2 - 2y_3^2 + \frac{3}{8}x_3^2. \end{aligned}$$

With the coordinates  $y_1, y_3, x_3$ , the quadratic form has a diagonal shape, which means that the basis associated with those coordinates is the polar basis of the form. If we want to express the basis, we need to obtain the matrix which changes the basis from polar to standard. By definition of the change of basis matrix, its columns are the polar basis vectors. Either we express the old variables  $(x_1, x_2, x_3)$  by new variables  $(y_1, y_3, x_3)$ , or equivalently we express the new ones by the old ones (which is easier). In the latter case, we need to compute inverse matrix.

$y_1 = 2x_1 + y_2 + \frac{1}{2}x_3 = 2x_1 + (x_2 - x_1) + \frac{1}{2}x_3$  and  $y_3 = \frac{1}{2}y_2 - \frac{1}{2}x_3 = -\frac{1}{2}x_1 + \frac{1}{2}x_3 - \frac{1}{2}x_3$ . The matrix for changing the basis from the polar basis to the standard basis is

$$T = \begin{pmatrix} 2 & 1 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

The inverse matrix is

$$T^{-1} = \begin{pmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{1}{2} \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence one of the polar bases of the given quadratic forms is (see the columns of the matrix),  $\{(1/3, 1/3, 0), (-2/3, 4/3, 0), (-1/2, 1/2, 1)\}$ .  $\square$

case of an equilateral triangle. Then this property holds for all triangles. Think about this argument!

**4.2.6. Transformations of quadrics.** After straight lines (which are mapped to straight lines again by all affine maps), the simplest objects in the analytic geometry of plane are the conic sections. These are given by quadratic equations in Cartesian coordinates. A conic is distinguished as a circle, ellipse, parabola or hyperbola, by examining the coefficients. There are two degenerate cases, namely a pair of lines or a point. We cannot distinguish a circle from an ellipse in affine geometry, but they are different in Euclidean geometry.

In analogy with the equations of conic sections in plane, we may discuss quadratic objects in Euclidean point spaces. These are defined in all orthonormal frames by quadratic equations, and they are known as *quadrics*. Consider a general quadratic equation for the coordinates  $(x_1, \dots, x_n)^T$  of a point  $A \in \mathcal{E}_n$

$$(1) \quad \sum_{i,j=1}^n a_{ij}x_i x_j + \sum_{i=1}^n a_i x_i + a = 0,$$

where it may be assumed by symmetry that  $a_{ij} = a_{ji}$  without loss of generality. This equation can be written as

$$f(u) + g(u) + a = 0$$

for a quadratic form  $f$  (i.e. the restriction of a symmetric bilinear form  $F$  to pairs of equal arguments), a linear form  $g$ , and a scalar  $a \in \mathbb{R}$ . We assume that at least one coefficient  $a_{ij}$  is nonzero. Otherwise the equation is linear and describes a hyperplane.

Notice that the equation (1) keeps the same shape under every affine coordinate transformation, i.e., it splits again into a nontrivial quadratic, linear and constant parts.

Recognizing which of the standard types of quadrics is determined by the given equation is an extremely useful tool (we shall see this later e.g. in multivariate analysis in Chapter 8) and thus we devote the next section to this topic.

Notice that the above observation is true for every fixed order of the multivariate polynomial expressions. In particular, there is the well defined concept of cubics, given by cubic equations and invariant with respect to all affine transformations. Similarly the quartics, quintics, etc.

We shall meet the cubic curves in plain in more details in Chapter 11 because of their fascinating use in cryptography.

### 3. Geometry of quadratic forms and quadrics

We shall start with the affine point of view and proceed to the Euclidean classification later.

**4.3.1. Linear transformations of quadratic forms.** Let us recall the bilinear symmetric forms  $F : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ , cf. 2.3.23, and the corresponding mappings  $f(x) = F(x, x)$  for all  $x \in \mathbb{R}^n$ .



**4.C.4.** Determine the type of conic section defined by

$$3x_1^2 - 3x_1x_2 + x_2 - 1 = 0.$$

**Solution.** Complete the squares:

$$\begin{aligned} 3x_1^2 - 3x_1x_2 + x_2 - 1 &= \frac{1}{3}(3x_1 - \frac{3}{2}x_2)^2 - \frac{3}{4}x_2^2 + x_2 - 1 \\ &= \frac{1}{3}y_1^2 - \frac{4}{3}(\frac{3}{4}x_2 - \frac{1}{2})^2 + \frac{1}{3} - 1 \\ &= \frac{1}{2}y_1^2 - \frac{4}{3}y_2^2 - \frac{2}{3}. \end{aligned}$$

According to the list 4.3.6, the given conic section is a hyperbola.  $\square$

**4.C.5.** By completing the squares, express the quadric

$$-x^2 + 3y^2 + z^2 + 6xy - 4z = 0$$

in such a way that one can determine its type from it.

**Solution.** Complete the square. Deal first with all terms involving an  $x$ . Obtain the equation

$$-(x - 3y)^2 + 9y^2 + 3y^2 + z^2 - 4z = 0.$$

There are no "unwanted" terms containing  $y$ , so repeat the procedure for  $z$ . This gives

$$-(x - 3y)^2 + 12y^2 + (z - 2)^2 - 4 = 0.$$

Conclude that there is a transformation of variables that leads to the equation (we can divide by 4 if desired)

$$-\bar{x}^2 + \bar{y}^2 + \bar{z}^2 - 1 = 0.$$

$\square$

We can tell the type of the conic section without transforming its equation to the form listed in 4.3.6. Every conic section can be expressed as

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33} = 0.$$

Determinants  $\Delta = \det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{32} & a_{33} \end{vmatrix}$  and

$\delta = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}$  are invariants of conic sections which means that they are not changed by Euclidean transformations (rotation and translation). Furthermore, the different types of conic sections have different signs of those determinants.

- $\Delta \neq 0$  for non-degenerate conic sections: ellipse for  $\delta > 0$ , hyperbola for  $\delta < 0$  and parabola for  $\delta = 0$

For a real ellipse (not imaginary), it is necessary that  $(a_{11} + a_{22})\Delta < 0$ .

- $\Delta = 0$  for degenerate conic sections, or pairs of lines.

The signs (or zero-value) of the determinants are really in-

variant to the coordinate transformation. Denote  $X = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$

and denote  $A$  as the matrix of the quadratic form. Then the corresponding conic section has equation  $X^TAX = 0$ . The

For an arbitrary basis  $\underline{e}$  on this vector space, the value  $f(x)$  on vector  $x = x_1e_1 + \dots + x_n e_n$  is given by the equation

$$(1) \quad f(x) = F(x, x) = \sum_{i,j} x_i x_j F(e_i, e_j) = x^T \cdot A \cdot x$$

where  $A = (a_{ij})$  is a symmetric matrix with elements  $a_{ij} = F(e_i, e_j)$ . We call such maps  $f$  *quadratic forms*, and the formula (1),  $f(x) = \sum_{i,j} a_{ij}x_i x_j$ , is called the *analytic formula* for the form  $f$ .

In general, by a quadratic form on vector space  $V$  is meant the restriction  $f(u)$  of a symmetric bilinear form  $F(u, v)$  to arguments of the type  $(u, u) \in V \times V$ . Evidently, the whole bilinear form  $F$  can be reconstructed from the values  $f(u)$  (unless we work over scalars  $\mathbb{Z}_2$ ) since

$$f(u + v) = F(u + v, u + v) = f(u) + f(v) + 2F(u, v).$$

This is called the *polarization process*.

Let us come back to the real vector spaces  $\mathbb{R}^n$ . If we change the basis  $\underline{e}$  to a different basis  $e'_1, \dots, e'_n$ , we get different coordinates  $x = S \cdot x'$  for the same vector (here  $S$  is the corresponding transformation matrix), and so

$$f(x) = (S \cdot x')^T \cdot A \cdot (S \cdot x') = (x')^T \cdot (S^T \cdot A \cdot S) \cdot x'.$$

Clearly, if we fix the second argument in the bilinear form  $F$ , we obtain a linear mapping  $V \rightarrow V^*$ , with coordinate description  $x \mapsto F(\cdot, x) = x^T \cdot A$ . Thus, the rank of the matrix  $A$  of the quadratic form  $f$  is the dimension of the image of this mapping and therefore it is independent of the choice of the coordinates. We call it the *rank of the quadratic form*  $f$ .

Our first task is to decide whether two quadratic forms can be transformed one into the other by a linear transformation.

We shall easily see that the matrix  $A$  of each quadratic form may be diagonal for a suitable choice of basis  $\underline{u}$  and this will resolve our task as well. In other words we request  $F(u_i, u_j) = 0$  for  $i \neq j$  for the corresponding symmetric bilinear form  $F$ . Each such basis is called the *polar basis* of the quadratic form  $f$ .

Later we shall see, we could even find a polar orthonormal basis (with respect to any scalar product). Nevertheless, without the use of the scalar product, there is a much simpler algorithm for finding a polar basis among all other bases. At the same time, this algorithm brings relevant information about the affine properties of the quadratic form. The algorithmic procedure in the proof of the next theorem is known as the *Lagrange algorithm*.

**Theorem.** Let  $f : V \rightarrow \mathbb{R}$  be a quadratic form on a real vector space  $V$  of dimension  $n$ . Then there exists a polar basis for  $f$  on  $V$ .

**PROOF.** (1) Let  $A$  be the matrix of  $f$  in basis  $\underline{u} = (u_1, \dots, u_n)$  on  $V$ , and assume  $a_{11} \neq 0$ . Then we may write

$$\begin{aligned} f(x_1, \dots, x_n) &= a_{11}x_1^2 + 2a_{12}x_1x_2 + \dots + a_{22}x_2^2 + \dots \\ &= a_{11}^{-1}(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n)^2 \\ &\quad + \text{terms not containing } x_1. \end{aligned}$$

standard form is obtained by rotation and translation. This is by a transformation to new coordinates  $x', y'$  satisfying

$$\begin{aligned} x &= x' \cos \alpha - y' \sin \alpha + c_1 \\ y &= x' \sin \alpha + y' \cos \alpha + c_2, \end{aligned}$$

or, in matrix form, for the new coordinates  $X' = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix}$ ,

$$(1) \quad X = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} \cos \alpha & -\sin \alpha & c_1 \\ \sin \alpha & \cos \alpha & c_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = MX'.$$

Put  $X = MX'$  into the conic section equation to obtain the equation in new coordinates

$$\begin{aligned} X^T A X &= 0 \\ (MX')^T A (MX') &= 0 \\ X'^T M^T A M X' &= 0. \end{aligned}$$

Denote by  $A'$  the matrix of the quadratic form in new coordinates. Then  $A' = M^T A M$ , where matrix

$$M = \begin{pmatrix} \cos \alpha & -\sin \alpha & c_1 \\ \sin \alpha & \cos \alpha & c_2 \\ 0 & 0 & 1 \end{pmatrix} \text{ has unit determinant, so}$$

$$\det A' = \det M^T \det A \det M = \det A = \Delta.$$

Necessarily, the determinant  $A_{33}$ , which is the algebraic complement of  $a_{33}$ , is invariant to the coordination transformation. For rotation only,  $\det A' = \det M^T \det A \det M$ .

In this case the matrix  $M = \begin{pmatrix} \cos \alpha & -\sin \alpha & 0 \\ \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$

and  $\det A'_{33} = \det A_{33} = \delta$ . For translation only,

$M = \begin{pmatrix} 1 & 0 & c_1 \\ 0 & 1 & c_2 \\ 0 & 0 & 1 \end{pmatrix}$  and this subdeterminant remains unchanged.

**4.C.6.** Determine the type of conic section

$$2x^2 - 2xy + 3y^2 - x + y - 1 = 0.$$

**Solution.** The determinant

$$\Delta = \begin{vmatrix} 2 & -1 & -\frac{1}{2} \\ -1 & 3 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -1 \end{vmatrix} = -\frac{23}{4} \neq 0,$$

hence it is a non-degenerate conic section. Moreover  $\delta = 5 > 0$ , therefore it is an ellipse. Furthermore  $(a_{11} + a_{22})\Delta = (2 + 3) \cdot (-\frac{23}{4}) < 0$ , so it is real ellipse.  $\square$

**4.C.7.** Determine the type of conic section  $x^2 - 4xy - 5y^2 + 2x + 4y + 3 = 0$ .

This suggest to transform the coordinates (i.e., change the basis)

$$x'_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n, \quad x'_2 = x_2, \dots, \quad x'_n = x_n.$$

This corresponds to the new basis

$$v_1 = a_{11}^{-1}u_1, \quad v_2 = u_2 - a_{11}^{-1}a_{12}u_1, \dots, \quad v_n = u_n - a_{11}^{-1}a_{1n}u_1$$

(As an exercise, compute the transformation matrix.) In the new basis the corresponding symmetric bilinear form satisfies  $F(v_i, v_i) = 0$  for all  $i > 1$  (compute it!). Thus  $f$  has the analytic form  $a_{11}^{-1}x_1'^2 + h$  in the new coordinates, where  $h$  is a quadratic form independent of the variable  $x_1$ .

It is often easier to choose  $v_1 = u_1$  in the new basis. Then  $f = f_1 + h$ , where  $f_1$  depends only on  $x'_1$ , while  $x'_1$  does not appear in  $h$ , but  $F(v_1, v_1) = a_{11}$ .

(2) Assume that after step (1),  $h$  is a matrix of rank less than  $n$  with a nonzero coefficient of  $x_2'^2$ . Then the same procedure can be repeated to obtain the expression  $f = f_1 + f_2 + h$ , where  $h$  contains only the variables with index greater than two. Proceed in this way until a diagonal form is obtained after  $n - 1$  steps, or in (say) the  $i$ -th step, the element  $a_{ii}$  is zero.

(3) If the last possibility occurs, and there exists some other element  $a_{jj} \neq 0$  with  $j > i$ , then it suffices to exchange the  $i$ -th and the  $j$ -th vector of the basis. Then continue according the previous procedure.

(4) Assume that the situation is  $a_{jj} = 0$  for all  $j \geq i$ . If there is no element  $a_{jk} \neq 0$  with  $j \geq i, k \geq i$ , then we are finished, since then the matrix is diagonal. If  $a_{jk} \neq 0$ , then we use the transformation  $v_j = u_j + u_k$  and we keep the other vectors of basis unchanged (i.e.  $x'_k = x_k - x_j$ , the other remain without change). Then  $h(v_j, v_j) = h(u_j, u_j) + h(u_k, u_k) + 2h(u_k, u_j) = 2a_{jk} \neq 0$  and we can continue as for case (1).  $\square$

**4.3.2. Affine classification of quadratic forms.**

The vectors in the basis obtained from the Lagrange algorithm can be rescaled by scalars such that the coefficients of the squares of variables are only the scalars 1, -1 and 0. Moreover, the following *law of inertia* says that the number of one's and minus one's does not depend on the choices in the course of the algorithm. These numbers are called the *signature of a quadratic form*. As before, there is a complete description of quadratic forms in the sense that two such forms may be transformed each one into the other by a linear transformation if and only if they have the same signature.

**Theorem.** For each nonzero quadratic form of rank  $r$  on a real vector space  $V$  there exists a natural number  $p$ , and  $r$  independent linear forms  $\varphi_1, \dots, \varphi_r \in V^*$  such that  $0 \leq p \leq r$  and  $f(u)$  becomes

$$(\varphi_1(u))^2 + \dots + (\varphi_p(u))^2 - (\varphi_{p+1}(u))^2 - \dots - (\varphi_r(u))^2.$$

Otherwise put, there exists a polar basis, in which  $f$  has an analytic formula

$$f(x_1, \dots, x_n) = x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_r^2.$$



**Solution.** The determinant  $\Delta = \begin{vmatrix} 1 & -2 & 1 \\ -2 & -5 & 2 \\ 1 & 2 & 3 \end{vmatrix} = -34 \neq 0$ ,

furthermore  $\delta = \begin{vmatrix} 1 & -2 \\ -2 & -5 \end{vmatrix} = -9 < 0$ , it is therefore a hyperbola.  $\square$

**4.C.8.** Determine the equation and type of conic section passing through the points

$$[-2, -4], [8, -4], [0, -2], [0, -6], [6, -2].$$

**Solution.** Input the coordinates of the points into the general conic section equation

$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_1x + a_2y + a = 0$$

There follows the linear equation system

$$\begin{aligned} 4a_{11} + 16a_{22} + 16a_{12} - 2a_1 - 4a_2 + a &= 0, \\ 64a_{11} + 16a_{22} - 64a_{12} + 8a_1 - 4a_2 + a &= 0, \\ 4a_{22} - 2a_2 + a &= 0, \\ 36a_{22} - 6a_2 + a &= 0, \\ 36a_{11} + 4a_{22} - 24a_{12} + 6a_1 - 2a_2 + a &= 0. \end{aligned}$$

In matrix form we perform operations

$$\begin{aligned} &\begin{pmatrix} 4 & 16 & 16 & -2 & -4 & 1 \\ 64 & 16 & -64 & 8 & -4 & 1 \\ 0 & 4 & 0 & 0 & -2 & 1 \\ 0 & 36 & 0 & 0 & -6 & 1 \\ 36 & 4 & -24 & 6 & -2 & 1 \end{pmatrix} \sim \dots \\ &\sim \begin{pmatrix} 4 & 16 & 16 & -2 & -4 & 1 \\ 0 & 4 & 0 & 0 & -2 & 1 \\ 0 & 0 & 64 & -8 & 12 & -9 \\ 0 & 0 & 0 & 24 & -36 & 27 \\ 0 & 0 & 0 & 0 & 3 & -2 \end{pmatrix} \sim \dots \\ &\sim \begin{pmatrix} 48 & 0 & 0 & 0 & 0 & -1 \\ 0 & 12 & 0 & 0 & 0 & -1 \\ 0 & 0 & 64 & 0 & 0 & 0 \\ 0 & 0 & 0 & 24 & 0 & 3 \\ 0 & 0 & 0 & 0 & 3 & -2 \end{pmatrix}. \end{aligned}$$

Then

$$a_{11} = 1, \quad a_{22} = 4, \quad a_{12} = 0, \quad a_1 = -6, \quad a_2 = 32.$$

The conic section has equation

$$x^2 + 4y^2 - 6x + 32y + 48 = 0.$$

Complete the terms  $x^2 - 6x$ ,  $4y^2 + 32y$  to squares. The result is

$$(x - 3)^2 + 4(y + 4)^2 - 25 = 0,$$

or rather

$$\frac{(x - 3)^2}{5^2} + \frac{(y + 4)^2}{\left(\frac{5}{2}\right)^2} - 1 = 0.$$

The conic section is an ellipse with centre at  $[3, -4]$ .  $\square$

The number  $p$  of positive diagonal coefficients in the matrix of the given quadratic form (and thus the number  $r - p$  of negative coefficients) does not depend on the choice of polar basis.

Two symmetric matrices  $A, B$  of dimension  $n$  are matrices of the same quadratic form in different bases if and only if they have the same rank and the same number of positive coefficients in the polar basis.

**PROOF.** By completing the squares,  $f(x_1, \dots, x_n) = \lambda_1 x_1^2 + \dots + \lambda_r x_r^2$ ,  $\lambda_i \neq 0$ , in a basis on  $V$ . Assume moreover that the first  $p$  coefficients  $\lambda_i$  are positive. Then the transformation  $y_1 = \sqrt{\lambda_1}x_1, \dots, y_p = \sqrt{\lambda_p}x_p, y_{p+1} = \sqrt{-\lambda_{p+1}}x_{p+1}, \dots, y_r = \sqrt{-\lambda_r}x_r, y_{r+1} = x_{r+1}, \dots, y_n = x_n$  yields the desired formula. The forms  $\varphi_i$  are exactly the forms from the dual basis in  $V^*$  to the obtained polar basis.

It remains to prove that  $p$  does not depend on the procedure. Assume that there is a formula for the same form  $f$  in the polar bases  $\underline{u}, \underline{v}$ , i.e.

$$\begin{aligned} f(x_1, \dots, x_n) &= x_1^2 + \dots + x_p^2 - x_{p+1}^2 - \dots - x_r^2 \\ f(y_1, \dots, y_n) &= y_1^2 + \dots + y_q^2 - y_{q+1}^2 - \dots - y_r^2. \end{aligned}$$

Denote the subspace generated by the first  $p$  vectors of the first basis by  $P = \langle u_1, \dots, u_p \rangle$ , and similarly  $Q = \langle v_{q+1}, \dots, v_r \rangle$ . Then for each  $0 \neq u \in P$ ,  $f(u) > 0$  while for  $0 \neq v \in Q$ ,  $f(v) \leq 0$ . Hence necessarily  $P \cap Q = \{0\}$ , and therefore  $\dim P + \dim Q \leq n$ . Hence  $p + (r - q) \leq r$ , so that  $p \leq q$ . By interchanging the subspaces,  $q \leq p$ , and so  $p = q$ .

Thus  $p$  is independent of the choice of the polar basis. Consequently for two matrices with the same rank and the same number of positive coefficients in the diagonal form of the corresponding quadratic form, the analytic formulas are the same.  $\square$

While discussing symmetric maps we talked about definite and semidefinite maps. The same discussion has an obvious meaning also for symmetric bilinear forms and quadratic forms.



#### TYPES OF QUADRATIC FORMS

A quadratic form  $f$  on a real vector space  $V$  is called

- (1) *positive definite* if  $f(u) > 0$  for all vectors  $u \neq 0$ ,
- (2) *positive semidefinite* if  $f(u) \geq 0$  for all vectors  $u \in V$ ,
- (3) *negative definite* if  $f(u) < 0$  for all vectors  $u \neq 0$ ,
- (4) *negative semidefinite* if  $f(u) \leq 0$  for all vectors  $u \in V$ ,
- (5) *indefinite* if  $f(u) > 0$  and  $f(v) < 0$  for two vectors  $u, v \in V$ .

The same names are used for symmetric matrices corresponding to quadratic forms. By the signature of a symmetric matrix is meant the signature of the corresponding quadratic form.

**4.C.9. Other characteristics and concepts of conic sections.** The axis of a conic section is a line of reflection symmetry for the conic section. From the canonical form of a conic section in polar basis (4.3.6) it can be shown that an ellipse and a hyperbola both have two axes ( $x = 0$  and  $y = 0$ ). A parabola has one axis ( $x = 0$ ). The intersection of a conic section and its axis is called a *conic section vertex*.

The numbers  $a, b$  from the canonical form of a conic section (which express the distance between vertices and the origin) are called the *length of semi-axes*. In the case of an ellipse and hyperbola, the axes intersect at the origin. This is a point of central symmetry for the conic section, called the *centre of the conic section*.

For practical problems involving conic sections, it is often easiest to describe them in parametric form. Often, this avoids contending with messy square roots.

Every point  $P$  on the parabola  $y^2 = 4ax$ ,  $a > 0$ , can be described by  $P = (x, y) = (at^2, 2at)$ , for real  $t$ . The standard parametric form for the parabola is the pair of equations

$$x = at^2 \quad y = 2at,$$

(Note that the roles of  $x$  and  $y$  are interchanged, so that the axis of symmetry is the line  $y = 0$ .) The tangent line at  $(at^2, 2at)$  has slope  $\frac{1}{t}$  and equation  $t(y - 2at) = (x - at^2)$ . The point  $F = (a, 0)$  on the axis is called the focus of the parabola, and the line  $x = -a$  is called the directrix. Each point on the parabola is equidistant from the focus and the directrix. This property can be used to define a parabola.

Every point  $P$  on the ellipse  $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$  can be described by  $P = (x, y) = (a \cos \theta, b \sin \theta)$ , where  $0 < b \leq a$ . The standard parametric form for the ellipse is the pair of equations

$$x = a \cos \theta, \quad y = b \sin \theta.$$

The tangent line at  $P$  has slope  $-\frac{b \cos \theta}{a \sin \theta}$  and consequently has equation  $(a \cos \theta)(y - b \sin \theta) = -b \cos \theta(x - a \cos \theta)$ . The positive number  $e$ , defined by  $b^2 = a^2(1 - e^2)$  is called the eccentricity of the ellipse. If  $e = 0$ , the ellipse becomes a circle or radius  $a = b$ . Otherwise  $0 < e < 1$ . The two points  $F_1 = (ae, 0)$  and  $F_2 = (-ae, 0)$  are the foci of the ellipse, and the lines  $x = \pm a/e$  are the directrices.

Every point  $P$  on the hyperbola  $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ ,  $0 < a, 0 < b$ , can be described by  $P = (x, y) = (a \cosh \theta, b \sinh \theta)$ . The standard parametric form for the hyperbola is the pair of equations

$$x = a \cosh \theta, \quad y = b \sinh \theta.$$

The tangent line at  $P$  has slope  $\frac{b \cosh \theta}{a \sinh \theta}$  and consequently has equation  $(a \cosh \theta)(y - b \sinh \theta) = b \cosh \theta(x - a \cosh \theta)$ . The positive number  $e$ , defined by  $b^2 = a^2(e^2 - 1)$  is called the eccentricity of the hyperbola. Necessarily,  $e > 1$ . The two points  $F_1 = (ae, 0)$  and  $F_2 = (-ae, 0)$  are the foci of the ellipse, and the lines  $x = \pm a/e$  are the directrices. A hyperbola has two asymptotes. In standard form, the equations are  $y = \pm(b/a)x$ .

**4.3.3. Theorem.** A symmetric real matrix  $A$  is positive definite if and only if all its leading principal minors are positive.

A symmetric real matrix  $A$  is negative definite if and only if  $(-1)^i |A_i| > 0$  for all leading principal submatrices  $A_i$ .

**PROOF.** The claim about negative definite forms follows immediately from the first part of the theorem. Just observe that  $A$  is positive definite if and only if  $-A$  is negative definite.

Suppose that the form  $f$  is positive definite. Then  $A = P^T E P = P^T P$  for a suitable regular matrix  $P$ . Hence  $|A| = |P|^2 > 0$ . Let  $\underline{u}$  be a chosen basis in which the form  $f$  has matrix  $A$ . The restrictions of  $f$  to the subspaces  $V_k = \langle u_1, \dots, u_k \rangle$  are positive definite forms  $f_k$  again, and the corresponding matrices in the bases  $u_1, \dots, u_k$  are the leading principal submatrices  $A_k$ . Thus  $|A_k| > 0$ , too.

In order to prove the other implication, analyse in detail the form of the transformations used in completing the square in the Lagrange algorithm. The transformation used in the first step always has an upper triangular matrix  $T$ . By rescaling, see the proof of Theorem 4.3.1, the matrix may have ones on the diagonal:



$$T = \begin{pmatrix} 1 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{n2}}{a_{11}} \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \end{pmatrix}.$$

Such a matrix of the transformation from basis  $\underline{u}$  to basis  $\underline{v}$  has several useful properties. In particular, its leading principal submatrices  $T_k$  formed by the first  $k$  rows and columns are the transformation matrices of a subspace  $P_k = \langle u_1, \dots, u_k \rangle$  from basis  $(u_1, \dots, u_k)$  to basis  $(v_1, \dots, v_k)$ . The leading principal submatrices  $A_k$  of the matrix  $A$  of the form  $f$  are matrices of restrictions of the form  $f$  to  $P_k$ . Therefore, the matrices  $A_k$  and  $A'_k$  of restrictions to  $P_k$  in basis  $\underline{u}$  and  $\underline{v}$  respectively satisfy  $A_k = T_k^T A'_k T_k$ , where  $T$  is the transformation matrix from  $\underline{u}$  to  $\underline{v}$ . The inverse matrix to an upper triangular matrix with one's on the diagonal is again an upper triangular matrix with one's on the diagonal. Hence we may similarly express  $A'$  in terms of  $A$ . Thus the determinants of the matrices  $A_k$  and  $A'_k$  are equal by Cauchy formula. We may conclude:

**Claim.** Let  $f$  be a quadratic form on  $V$ ,  $\dim V = n$ . Let  $\underline{u}$  be a basis of  $V$  such that the items (3) and (4) from the Lagrange algorithm while finding the polar basis are not needed. Then the analytic formula

$$f(x_1, \dots, x_n) = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \dots + \lambda_r x_r^2$$

is obtained where  $r$  is the rank of the form  $f$ ,  $\lambda_1, \dots, \lambda_r \neq 0$  and for the leading principal submatrices of the (former) matrix  $A$  of quadratic form  $f$ ,  $|A_k| = \lambda_1 \lambda_2 \dots \lambda_k$ ,  $k \leq r$ .

**4.C.10. Existence of foci.** For an ellipse with lengths of semi-axes  $a > b$ , show that the sum of the distances from any point on the ellipse to the two foci is constant, namely  $2a$ .

**Solution.** If  $P = (a \cos \theta, b \sin \theta)$  and  $F_1 = (ae, 0)$ , then

$$\begin{aligned} |PF_1|^2 &= (a \cos \theta - ae)^2 + b^2 \sin^2 \theta \\ &= a^2 \cos^2 \theta - 2a^2 e \cos \theta + a^2 e^2 + a^2(1 - e^2) \sin^2 \theta \\ &= a^2[-2e \cos \theta + e^2 - e^2(1 - \cos^2 \theta)] \\ &= a^2(1 - e \cos \theta)^2. \end{aligned}$$

So  $|PF_1| = a(1 - e \cos \theta)$ . Similarly  $|PF_2| = a(1 + e \cos \theta)$ . Hence  $|PF_1| + |PF_2| = 2a$ .  $\square$

**Solution.** (Alternative). Consider the points  $X = [x, y]$ , which satisfy the property  $|F_1X| + |F_2X| = 2a$ . Coordinate-wise, this implies the equation

$$\sqrt{(x + ae)^2 + y^2} + \sqrt{(x - ae)^2 + y^2} = 2a$$

Rewrite this as

$$\sqrt{(x + ae)^2 + y^2} = 2a - \sqrt{(x - ae)^2 + y^2}$$

Square, simplify and square again to get

$$(1 - e^2)x^2 + y^2 = a^2(1 - e^2).$$

Substitute  $b^2 = a^2(1 - e^2)$  and divide by  $b^2$  to obtain

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

which is the ellipse in standard form.  $\square$

**Remark.**

Similarly, the *hyperbola foci* are the points  $F_1, F_2$ , which satisfy  $||F_2X| - |F_1X|| = 2a$  for an arbitrary  $X$  on the hyperbola. You can check this in same the way as above for the ellipse, with  $F_1 = [ae, 0], F_2 = [-ae, 0], ae = \sqrt{a^2 + b^2}$ .

*Parabola focus* If the parabola has equation  $x^2 = 2py$ , the focus is the point  $F$  with coordinates  $F = [0, \frac{p}{2}]$ . It is characterized by the fact that the distance between this point and an arbitrary  $X$  on parabola is equal to the distance between  $X$  and line  $y = -\frac{p}{2}$ .

**4.C.11.** Find the foci of the ellipse  $x^2 + 2y^2 = 2$ .

**Solution.** From the equation that semi-axes lengths are  $a = \sqrt{2}$  and  $b = 1$ . Compute (see 4.C.10):  $ae = \sqrt{a^2 - b^2} = 1$  The foci coordinates are at  $[-1, 0]$  and  $[1, 0]$ .  $\square$

**4.C.12.** Prove that the product of the distances between the foci of an ellipse and any tangent line is constant. Find the value of the constant.

**Solution.** Every point  $T$  on the ellipse has coordinates  $T = (x, y)$  where  $x = a \cos \theta, y = b \sin \theta$  for some  $\theta$ . The tangent line to the ellipse at  $T$  has equation

$$\begin{aligned} y - b \sin \theta &= -\frac{b \cos \theta}{a \sin \theta}(x - a \cos \theta). \\ a(\sin \theta)(y - b \sin \theta) &= -b(\cos \theta)(x - a \cos \theta). \end{aligned}$$

After each step in this procedure, the resulting matrix contains zeros under the diagonal in the already processed columns. At the same time, all the principal minors remain the same.

Consequently if the leading principal minors are nonzero, then the next diagonal term in  $A$  is nonzero and we do not need other steps than completing the squares. Moreover,  $\lambda_i = |A_i|/|A_{i-1}|$ . This proves the following:

**Corollary (Jacobi theorem).** Let  $f$  be a quadratic form of rank  $r$  on a vector space  $V$  with matrix  $A$  for the basis  $\underline{u}$ . Steps other than completing the square are not required if and only if the leading principal submatrices of  $A$  satisfy  $|A_1| \neq 0, \dots, |A_r| \neq 0$ . Then there exists a polar basis in which  $f$  has the analytic formula

$$f(x_1, \dots, x_n) = |A_1|x_1^2 + \frac{|A_2|}{|A_1|}x_2^2 + \dots + \frac{|A_r|}{|A_{r-1}|}x_r^2.$$

Hence if all leading principal minors are positive, then  $f$  is positive definite by the Jacobi theorem and the Sylvester criterion is proved.  $\square$

**4.3.4. Euclidean classification of quadratic forms.** We come back to the discussion of equation 4.2.6(1) in the Euclidean context.



Begin with its quadratic part, i.e. bilinear symmetric form  $F : V \times V \rightarrow \mathbb{R}$ .

Assume now that the real vector space is equipped with a scalar product and choose an orthogonal basis. Then the matrix of the bilinear form  $F$ , which is the same as the matrix of  $f$ , transforms under a change of coordinates in such a way that for orthogonal changes it coincides with the transformation of a matrix of a linear map (i.e., then  $S^{-1} = S^T$ ). This result can be interpreted as the following observation:

**Proposition.** Let  $V$  be a real vector space with a scalar product. Then formula

$$\varphi \mapsto F, \quad F(u, u) = \langle \varphi(u), u \rangle$$

defines a bijection between symmetric linear maps and quadratic forms on  $V$ .

**PROOF.** Each bilinear form with a fixed second argument becomes a linear form  $F(\cdot, u)$ . In the presence of a scalar product, it is given by the formula  $F(v, u) = v \cdot w$  for a suitable vector  $w$ . Put  $\varphi(u) = w$ . Directly from the coordinate expression 4.3.1(1) displayed above,  $\varphi$  is the linear map with symmetric matrix  $A$ . Hence it is selfadjoint.

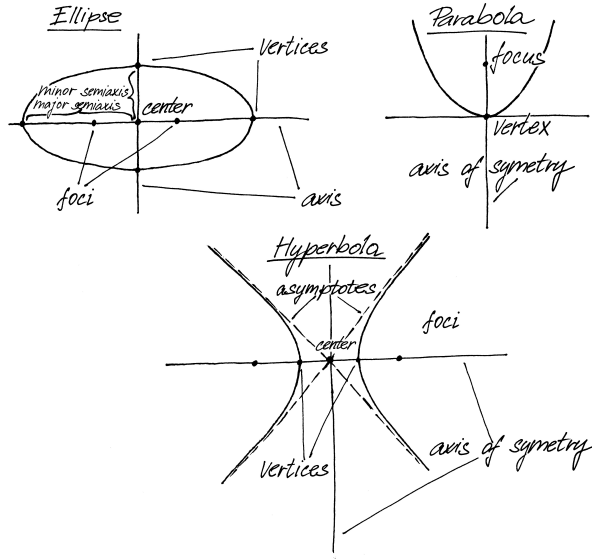


On the other hand, each symmetric map  $\varphi$  defines a symmetric bilinear form  $F$  by formula  $F(u, v) = \langle \varphi(u), v \rangle = \langle u, \varphi(v) \rangle$ , and thus is also a quadratic form.  $\square$

It is immediate that for each quadratic form  $f$  there exists an orthonormal basis in which  $f$  has a diagonal matrix. The values on the diagonal are the eigenvalues of the matrix and they are determined uniquely up to their order.

The rank of  $f$  equals the dimension of the image of the corresponding map  $\varphi$ .





This meets the  $x$ -axis at the point  $(a/\cos\theta, 0)$ . The distance from the focus  $F_1$  to the tangent line is  $D_1 = [a/\cos\theta - ae]\sin\varphi$  where  $\tan\varphi = \pm \frac{b\cos\theta}{a\sin\theta}$ . Eliminate  $\varphi$  to get

$$\begin{aligned} D_1^2 &= a^2(1 - e\cos\theta)^2 \left[ \frac{b^2}{a^2\sin^2\theta + b^2\cos^2\theta} \right] \\ &= (1 - e\cos\theta)^2 \left[ \frac{b^2}{\sin^2\theta + (1 - e^2)\cos^2\theta} \right] \\ &= (1 - e\cos\theta)^2 \left[ \frac{b^2}{1 - e^2\cos^2\theta} \right] \\ &= b^2(1 - e\cos\theta) \left[ \frac{1}{1 + e\cos\theta} \right] \end{aligned}$$

Since  $D_2$  is the same as  $D_1$  with  $e$  replaced by  $-e$ , it follows that  $D_1D_2 = b^2$ .  $\square$

**Solution.** (Alternative). Consider the polar basis. The ellipse matrix has diagonal shape  $\text{diag}(\frac{1}{a^2}, \frac{1}{b^2}, -1)$  and the tangent equation at  $X=[x_0, y_0]$  is  $\frac{x_0}{a^2}x + \frac{y_0}{b^2}y = 1$ . The distance between  $F_1, F_2 = [\mp ae, 0]$  and this line equals

$$\frac{1 \pm ae \frac{x_0}{a^2}}{\sqrt{\frac{x_0^2}{a^4} + \frac{y_0^2}{b^4}}}$$

Its product is

$$\frac{1 - e^2 \frac{x_0^2}{a^4}}{\frac{x_0^2}{a^4} + \frac{y_0^2}{b^4}}$$

If we substitute  $a^2e^2 = a^2 - b^2$  and  $\frac{y_0^2}{b^2} = 1 - \frac{x_0^2}{a^2}$  (the point  $X$  is lying on the ellipse), we find that the previous term equals  $b^2$ .  $\square$

**4.C.13. Projective approach to conic section.** Projective space gives an ability to approach the conic section from a new perspective (compare with 4.4.11). We can understand conic sections in  $\mathcal{E}_2$  defined by the quadratic form

$$f(x, y) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{23}y + a_{33}$$

**4.3.5. Classification of quadrics.** We return to the equation 4.2.6(1). The above results enable us to rewrite this equation in suitable orthonormal frame of the difference space as

$$\sum_{i=1}^n \lambda_i x_i^2 + \sum_{i=1}^n b_i x_i + b = 0.$$

Hence we may assume that the quadric is given in this form.

In the next step, we “complete the square” for the coordinates  $x_i$  with  $\lambda_i \neq 0$ , which “absorbs” the squares together with the linear terms in the same variable. So only linear terms are left corresponding to variables for which the coefficient of the quadratic term is zero. We have

$$\sum_{i=1}^n \lambda_i (x_i - p_i)^2 + \sum_j b_j x_j + c = 0.$$

where the summation over  $j$  is only for  $j$  satisfying  $\lambda_j = 0$ .

This corresponds to a translation of the origin about the vector with coordinates  $p_i$ . To such a choice of basis of the difference space the desired diagonal form is in the quadratic part. In the identification of quadratic forms with linear maps derived above, this means that  $\varphi$  is diagonal on the orthogonal complement of its kernel. If there are also some linear terms left, the orthonormal basis of the difference space can be adjusted for the kernel of  $\varphi$  so that the corresponding linear form is a multiple of the first term of the dual basis. Hence the final formula reads

$$\sum_{i=1}^k \lambda_i y_i^2 + by_{k+1} + c = 0,$$

where  $k$  is the rank of matrix of quadratic form  $f$ . If  $b \neq 0$ , it can be arranged that the constant  $c$  in the equation is zero by a further change of the origin.

Hence the linear term may (but does not have to) appear only in the case that the rank of  $f$  is less than  $n$ , while  $c \in \mathbb{R}$  may be nonzero only if  $b = 0$ . The resulting equations are called the *canonical analytic formulas* for quadrics.

**4.3.6. The case of  $\mathcal{E}_2$ .** As an example of the previous procedure, we discuss the simplest case of a non-trivial dimension, namely dimension two. The original equation has the form

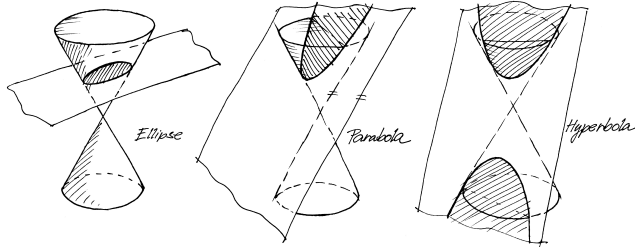


$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_1x + a_2y + a = 0.$$

By a suitable choice of a basis of the difference space, and the subsequent completion of the square, it is written in the form (using the same notation  $x, y$  for the new coordinates):

$$a_{11}x^2 + a_{22}y^2 + a_1x + a_2y + a = 0$$

where  $a_i$  is nonzero only in the case that  $a_{ii}$  is zero. By the last step of the general procedure, exactly one of the following equations is involved (notice the equations of quadrics are determined up to a multiple):



$0 = x^2/a^2 + y^2/b^2 + 1$	empty set
$0 = x^2/a^2 + y^2/b^2 - 1$	ellipse
$0 = x^2/a^2 - y^2/b^2 - 1$	hyperbola
$0 = x^2/a^2 - 2py$	parabola
$0 = x^2/a^2 + y^2/b^2$	point
$0 = x^2/a^2 - y^2/b^2$	2 concurrent lines
$0 = x^2 - a^2$	2 parallel lines
$0 = x^2$	2 identical lines
$0 = x^2 + a^2$	empty set

as a set of points in projective plane  $\mathcal{P}^2$  with homogenous coordinates  $(x : y : z)$ , which are the zero points of the homogenous quadratic form

$$f(x, y, z) = a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}xz + 2a_{23}yz + a_{33}z^2.$$

Or rather  $f(v) = v^T Av$ , where  $v$  is a column vector with coordinates  $(x, y, z)$  and matrix  $A$  is symmetric matrix  $(a_{ij})$ . By theorem 4.3.2, there exists a basis in which this quadratic form has one of the following equations

$$f(x, y, z) = x^2 + y^2 + z^2, \quad f(x, y, z) = x^2 + y^2 - z^2.$$

In the former case there is only one solution of  $f(x, y, z) = 0$  and therefore the original form does not represent a real conic section. The second quadratic form represents a cone in  $\mathbb{R}^3$ . We obtain the corresponding conic section by moving back to inhomogeneous coordinates. That means intersecting the cone with the plane which has the equation  $z = 1$  in the original basis. Immediately we obtain the conic section classification from 4.29., which corresponds to the intersecting cone in  $\mathbb{R}^3$  with different planes. Non-degenerate sections are depicted. Degenerate sections are those which pass through the vertex of the cone.

We define the following useful terms for a conic section in projective plane :

Points  $P, Q \in \mathcal{P}^2$  corresponding to one-dimensional subspaces  $\langle p \rangle, \langle q \rangle$  (generated by vectors  $p, q \in \mathbb{R}^3$ ) are called *polar conjugated* with respect to conic section  $f$ , if  $F(p, q) = 0$ , or rather  $p^T Aq = 0$ .

Point  $P = \langle p \rangle$  is called *singular point* of conic section  $f$ , when it is polar conjugated with respect to  $f$  with all points of the plane, so  $F(p, x) = 0 \quad \forall x \in \mathcal{P}^2$ . In other words,  $Ap = 0$ . Hence the matrix  $A$  of the conic section does not have maximal rank and therefore defines a degenerate conic section. Non-degenerate conic sections do not contain singular points.

The set of all points  $X = \langle x \rangle$  are called polar conjugated with  $P = \langle p \rangle$  polar of the point  $P$  with respect to the conic section  $f$ . It is therefore the set of points for which  $F(p, x) = p^T Ax = 0$ . Because the polar is given by a linear combination of coordinates, it is always (in the non-singular case) a line. The following explains the geometric interpretation of polar.

**4.C.14. Polar characterization.** Consider a non-degenerate conic section  $f$ . The polar of a point  $P \in f$  with respect to  $f$  is the tangent to  $f$  with the touch point  $P$ . The polar of

The origin of the Cartesian coordinates is the *center* of the studied conic. The new orthonormal basis of the difference space gives the direction of *semiaxes*. The final coefficients  $a, b$  then give the lengths of the semiaxes in the non-degenerate directions.

#### 4. Projective geometry

In many elementary texts on analytic geometry, the authors finish with the affine and Euclidean objects described above. The affine and Euclidean geometries are sufficient for many practical problems, but not for all problems.

For instance in processing an image from a camera, angles are not preserved and parallel lines may (but do not have to) intersect.

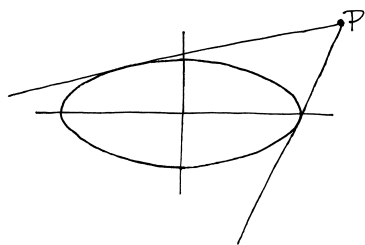
Moreover, it is often difficult to distinguish very small angles from zero angles, and thus it would be convenient to have tools which do not need such distinguishing.

The basic idea of projective geometry is to extend affine spaces by points at infinity. This permits an easy way to deal with linear objects such as points, lines, planes, projections, etc.

**4.4.1. Projective extension of affine plane.** We begin with the simplest interesting case, namely geometry in a plane. If we imagine the points in the plane  $\mathcal{A}_2$  as the plane  $z = 1$  in  $\mathbb{R}^3$ , then each point  $P$  in the affine plane is represented by a vector  $u = (x, y, 1) \in \mathbb{R}^3$ . So it is represented also by a one-dimensional subspace  $\langle u \rangle \subset \mathbb{R}^3$ . On the other hand, almost every one-dimensional subspace in  $\mathbb{R}^3$  intersects the plane in exactly one point  $P$ . The vectors of such a subspace are given by coordinates  $(x, y, z)$  uniquely up to a common scalar multiple. Only the subspaces corresponding to vectors  $(x, y, 0)$  do not intersect the plane.

#### PROJECTIVE PLANE

**Definition.** The projective plane  $\mathcal{P}_2$  is the set of all one-dimensional subspaces in  $\mathbb{R}^3$ . The homogeneous coordinates of a point  $P = (x : y : z)$  in the projective plane are triples of real numbers given up to a common scalar multiple, at least one of which must be nonzero. A straight line in the projective plane is defined as a set of one-dimensional subspaces (i.e. points in  $\mathcal{P}_2$ ) which generate a two-dimensional subspace (i.e. a plane) in  $\mathbb{R}^3$ .



the point  $P \notin f$  is the line defined by the touch points of the tangents to  $f$  passing through  $P$ .

**Solution.** First consider  $P \in f$ . Suppose that the polar of  $P$ , defined by  $F(p, x) = 0$ , intersects  $f$  in  $Q = \langle q \rangle \neq P$ . Then  $F(p, q) = 0$  and  $f(q) = F(q, q) = 0$ . For an arbitrary point  $X = \langle x \rangle$  lying on  $P$  and  $Q$ ,  $x = \alpha p + \beta q$  for some  $\alpha, \beta \in \mathbb{R}$ . Because of the bilinearity and symmetry of  $F$ ,

$$f(x) = F(x, x) = \alpha^2 F(p, p) + 2\alpha\beta F(p, q) + \beta^2 F(q, q) = 0.$$

So every point  $X$  of the line lies on the conic section  $f$ . However, when the conic section contains a line, it has to be degenerate, which is a contradiction.

The claim for  $P \notin f$  follows from the corollary of the symmetry of the bilinear form  $F$ . When the  $Q$  lies on the polar of  $P$ , then  $P$  lies on the polar of  $Q$ .

□

Using polar conjugates we can find the axes and the centre of the conic sections without using the Lagrange algorithm.

Consider the conic section matrix as a block matrix

$$A = \begin{pmatrix} \bar{A} & a \\ a^T & \alpha \end{pmatrix},$$

where  $\bar{A} = (a_{ij})$  for  $i, j = 1, 2$ ,  $a$  is vector  $(a_{13}, a_{23})$  and  $\alpha = a_{33}$ . This means that the conic section is defined by the equation

$$u^T \bar{A} u + 2a^T u + \alpha = 0$$

for a vector  $u = (x, y)$ . Now we show that

**4.C.15.** The axes of a conic section are the polars of the points at infinity determined by the eigenvectors of the matrix  $\bar{A}$ .

**Solution.** Because of the symmetry of  $\bar{A}$  in the basis of its eigenvectors, it has a diagonal shape  $D = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$ , where  $\lambda, \mu \in \mathbb{R}$  and this basis is orthogonal. Denote by  $U$  the matrix changing basis to a basis of eigenvectors (columns), then the conic section matrix is

$$\begin{pmatrix} U^T & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{A} & a \\ a^T & \alpha \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} D & U^T a \\ a^T U & \alpha \end{pmatrix}$$

For a concrete example, consider two parallel lines in the affine plane  $\mathbb{R}^2$

$$L_1 : y - x - 1 = 0, \quad L_2 : y - x + 1 = 0.$$

If the points of lines  $L_1$  and  $L_2$  are finite points in projective space  $\mathcal{P}_2$ , then their homogeneous coordinates  $(x : y : z)$  satisfy equations

$$L_1 : y - x - z = 0, \quad L_2 : y - x + z = 0.$$

the intersection  $L_1 \cap L_2$  is the point  $(1 : 1 : 0) \in \mathcal{P}_2$  in this context. It is the point at infinity corresponding to the common direction vector of the lines.

**4.4.2. Affine coordinates in the projective plane.** If we begin with the projective plane and if we want to see the affine plane as its “finite” part, then instead of the plane  $z = 1$  we may take another plane  $\sigma$  in  $\mathbb{R}^3$  which does not pass through the origin  $0 \in \mathbb{R}^3$ . Then the finite points are those one-dimensional subspaces which have a nonempty intersection with the plane  $\sigma$ .

Consider the two parallel lines from the previous paragraph. Let us choose the plane  $y = 1$  to obtain two lines in the affine plane

$$L'_1 : 1 - x - z = 0, \quad L'_2 : 1 - x + z = 0$$

The “infinite” points of the former affine plane are given by  $z = 0$ . The lines  $L'_1$  and  $L'_2$  intersect at the “finite” point  $(x, z) = (1, 0)$ . This corresponds to the geometric concept that two parallel lines  $L_1, L_2$  in the affine plane meet at infinity, at the point  $(1 : 1 : 0)$ , but this point becomes finite in different finite affine plane.

**4.4.3. Projective spaces.** We shall not go for an axiomatic definition of projective spaces here. We generalize the procedure in the affine plane for each finite dimension instead.

By choosing an arbitrary affine hyperplane  $\mathcal{A}_n$  in the vector space  $\mathbb{R}^{n+1}$  which does not pass through origin, we may identify the points  $P \in \mathcal{A}_n$  with one-dimensional subspaces generated by these points. The remaining one-dimensional subspaces determine a hyperplane parallel to  $\mathcal{A}_n$ . They are called *infinite points* in the projective extension  $\mathcal{P}_n$  of the affine plane  $\mathcal{A}_n$ .

The set of infinite points in  $\mathcal{P}_n$  is always a projective space of dimension one less. An affine straight line has only one infinite point in its projective extension (both ends of the line “intersect” at infinity and thus the projective line looks like a circle). The projective plane has a projective line of infinite points, the three-dimensional projective space has a projective plane of infinite points etc.

More generally, we can define the *projectivization of a vector space*. For an arbitrary vector space  $V$  of dimension  $n + 1$ , we define

$$\mathcal{P}(V) = \{P \subset V; P \subset V, \dim V = 1\}.$$



So in this basis there is the canonical form defined by vector  $U^T a$  (up to a translation). Specifically, denote the eigenvectors by  $v_\lambda, v_\mu$ , and then

$$\lambda(x + \frac{a^T v_\lambda}{\lambda})^2 + \mu(y + \frac{a^T v_\mu}{\mu})^2 = \frac{(a^T v_\lambda)^2}{\lambda} + \frac{(a^T v_\mu)^2}{\mu} - \alpha.$$

This means that the eigenvectors are the direction vectors of the conic section axes (main directions). The axes equations in this basis are  $x = -\frac{a^T v_\lambda}{\lambda}$  and  $y = -\frac{a^T v_\mu}{\mu}$ . The axes coordinates  $u_\lambda$  and  $u_\mu$  in the standard basis satisfy  $v_\lambda^T u_\lambda = -\frac{a^T v_\lambda}{\lambda}$  and  $v_\mu^T u_\mu = -\frac{a^T v_\mu}{\mu}$ , because  $v_\lambda^T(\lambda u_\lambda + a) = 0$  and  $v_\mu^T(\mu u_\mu + a) = 0$ . These equations are equivalent to the equations  $v_\lambda^T(\bar{A}u_\lambda + a) = 0$  and  $v_\mu^T(\bar{A}u_\mu + a) = 0$  which are the polar equations of the points defined by the vectors  $v_\lambda$  a  $v_\mu$ .  $\square$

**4.C.16. Remark.** A corollary of the previous claim is that the centre of the conic section is polar conjugated with all points at infinity. The coordinates of the centre  $s$  then satisfy the equation  $\bar{A}s + a = 0$ .

If  $\det(A) \neq 0$ , then the equation  $\bar{A}s + a = 0$  for centre coordinates has exactly one solution if  $\delta = \det(\bar{A}) \neq 0$ , and no solutions if  $\delta = 0$ . That means that, regarding non-degenerate conic sections, the ellipse and the hyperbola have exactly one centre. The parabola has no centre. (its centre is point at infinity).

**4.C.17.** Prove that the angle between the tangent to the parabola (with arbitrary touch point) and the parabola axis is the same as the angle between the tangent and the line connecting the focus and the point of tangency

**Solution.** The polar (i.e. tangent) of a point  $X=[x_0, y_0]$  to a parabola defined by the canonical equation in the polar basis is a line satisfying

$$(x_0, y_0, 1) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -p \\ 0 & -p & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = x_0x - py - py_0 = 0$$

The cosine of the angle between the tangent and the axis of the parabola ( $x = 0$ ) is given by the dot product of the corresponding unit direction vectors. The unit direction vector of the tangent is  $\frac{1}{\sqrt{p^2+x_0^2}}(p, x_0)$  and therefore

$$\frac{1}{\sqrt{p^2+x_0^2}}(p, x_0) \cdot (0, 1) = \frac{x_0}{\sqrt{p^2+x_0^2}}$$

Now we show that this is the same as the cosine of the angle between the tangent and line connecting the focus  $F=[0, \frac{p}{2}]$ , and the touch point  $X$ . The unit direction vector of the connecting line is

$$\frac{1}{\sqrt{x_0^2 + (y_0 - \frac{p}{2})^2}}(x_0, y_0 - \frac{p}{2}).$$

For the cosine of the angle,

$$\frac{1}{\sqrt{p^2+x_0^2}} \frac{1}{\sqrt{x_0^2 + (y_0 - \frac{p}{2})^2}}(x_0y_0 + \frac{px_0}{2})$$

By choosing a basis  $\underline{u}$  in  $V$  we obtain *homogeneous coordinates* on  $\mathcal{P}(V)$ . For a  $P \in \mathcal{P}(V)$  we use the nonzero vector  $u \in V$  and the coordinates of this vector in a basis  $\underline{u}$ . The points of the projective space  $\mathcal{P}(V)$  are called *geometric points*. Their generators in  $V$  are called *arithmetic representatives*.

In the chosen homogeneous coordinates, we fix one of them to be one. Thus we exclude all points of the projective space  $\mathcal{P}(V)$  which have this coordinate equal to zero. We have an embedding of  $n$ -dimensional affine space  $\mathcal{A}_n \subset \mathcal{P}(V)$ . This is precisely the construction used in the example on the projective plane.

**4.4.4. Perspective projection.** Our basic “projective” concepts are now (projective) points, lines, planes, etc., together with their incidencies (i.e. point in a line, line in a plane, etc.). Thus the morphisms of the projective geometry must respect these.

The best known example is given by perspective projections  $\mathbb{R}^3 \rightarrow \mathbb{R}^2$ . Imagine that an observer sitting in the origin observes “one half of the world”, that is, the points  $(X, Y, Z) \in \mathbb{R}^3$  with  $Z > 0$ . The observer sees the image “projected” on the screen given by plane  $Z = f > 0$ .



picture missing!

Thus a point  $(X, Y, Z)$  in the “real world” projects to a point  $(x, y)$  on the screen as follows:

$$x = -f \frac{X}{Z}, \quad y = -f \frac{Y}{Z}.$$

This is a nonlinear formula. The accuracy of calculations are problematic when  $Z$  is small.

By extending this transformation to a map between projective spaces, we have  $(X : Y : Z : W) \mapsto (x : y : z) = (-fX : -fY : Z)$ , which is well defined for all  $Z > 0$ . That is, a map described by a simple linear formula

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \\ W \end{pmatrix}$$

This simple expression defines the perspective projection for finite points in the open half-space in  $\mathbb{R}^3 \subset \mathcal{P}_3$  which we

Substitute  $y_0 = \frac{x_0^2}{2p}$  to obtain  $\frac{x_0}{\sqrt{p^2+x_0^2}}$ .

This example shows that lightrays striking parallel with axis of parabolic mirror are reflecting to the focus and, vice versa, light rays going through focus reflect in direction parallel with axis of parabola. This is the principle of many devices such as parabolic reflectors.  $\square$

**Solution.** (Alternative) At the point  $P = (at^2, 2at)$  on the parabola, the tangent line has slope  $(1/t)$  and the focus is at  $(a, 0)$ . So the line joining  $P$  to the focus  $F$  has slope  $\frac{2at-0}{at^2-a} = \frac{2t}{t^2-1}$ . If  $\theta$  is the angle between the tangent line and the  $x$ -axis, then  $\tan \theta = 1/t$ , so

$$\tan 2\theta = \frac{2 \tan \theta}{1 - \tan^2 \theta} = \frac{2/t}{(1 - 1/t^2)} = \frac{2t}{t^2 - 1}$$

By subtraction, the angle between the tangent line and the line joining  $P$  to the focus is  $\theta$ .

Note that the tangent line meets the  $x$ -axis at  $Q$  where  $Q = (-at^2, 0)$ . The result follows from showing that  $|FP| = |FQ|$ , and hence the triangle  $QFP$  is isosceles.  $\square$

You can find many more examples on quadrics on 4.D.1

substitute as points with  $W = 1$ . In this way we eliminate problems with points whose image runs to infinity. Indeed, if the  $Z$ -coordinate of a real point is close to zero, then the value of the third homogeneous coordinate of the image is close to zero, i.e. it corresponds to a point close to infinity.

**4.4.5. Affine and projective transformations.** Each in-



jective linear map  $\varphi : V_1 \rightarrow V_2$  between vector spaces maps one-dimensional subspaces to one-dimensional subspaces. Therefore, we have a map on projectivizations  $T : \mathcal{P}(V_1) \rightarrow \mathcal{P}(V_2)$ . We shall call such maps *projective maps* or *homographies*. A slightly more general concept posits the property that a map  $C : \mathcal{P}(V_1) \rightarrow \mathcal{P}(V_2)$  is bijective and maps projective lines to projective lines. They are called *collineations*. Of course, every invertible homography is a collineation.<sup>1</sup>

Otherwise put, the projective map is a map between projective spaces such that in each system of homogeneous coordinates on the domain and image, it is given by the multiplication by a matrix. More generally if the auxiliary linear map is not injective, then we need to define the projective map only outside of its kernel, that is, on points whose homogeneous coordinates do not map to zero. This is, what we saw in the previous paragraph when discussing the perspective projections.

In general, each collineation does not have to be a homography. For example, if we work with complex vector spaces, then we may use the field homomorphism  $z \mapsto \bar{z}$ , i.e. the conjugation, to define the map  $F : (z_0 : \cdots : z_n) \mapsto (\bar{z}_0 : \cdots : \bar{z}_n)$  in any fixed homogeneous coordinates. Clearly this is bijective and since the colinearity is computed by subdeterminants (and the determinants are equivariant with respect to each field homomorphism) the colinearity is preserved. Thus,  $F$  is a collineation which does not come from a linear automorphism of the vector space  $\mathbb{C}^{n+1}$ . Fortunately, there is the so called *fundamental theorem of projective geometry* saying that in dimensions at least two, each collineation is composed of a homography and an collineation coming from a field homomorphism in the above way. Of course, in dimension one, every bijection is a collineation and this concept does not make much sense there.

Notice, there are no nontrivial field homomorphisms on the real scalars  $\mathbb{R}$  and thus the two concepts coincide for our real projective spaces in dimensions at least two.

Since injective maps  $V \rightarrow V$  of a vector space to itself are invertible, all projective maps of projective space  $\mathcal{P}_n$  to itself are invertible. They are also called *regular collineations* or *projective transformations*. In homogeneous coordinates, they correspond to invertible matrices of dimension  $n + 1$ . Two such matrices define the same projective transformation if and only if one is a (nonzero) multiple of the other.

<sup>1</sup>Since projective geometry is an old and rich mathematical discipline with very abstract current versions, there is a lot of terminology and different names for similar or same objects involved.

If we choose the first coordinate as the one whose vanishing defines infinite points, then the transformations preserving infinite points are given by matrices whose first row vanishes up to its first element. If we wish to switch to affine coordinates of finite points, (i.e the first coordinate is fixed at one), the first element in the first row also equals one. Hence the matrices of collineations preserving finite points of the affine space have the form:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ b_1 & a_{11} & \cdots & a_{1n} \\ \vdots & & \ddots & \\ b_n & a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

where  $b = (b_1, \dots, b_n)^T \in \mathbb{R}^n$  and  $A = (a_{ij})$  is an invertible matrix of dimension  $n$ . The action of such a matrix on the vector  $(1, x_1, \dots, x_n)$  is exactly a general affine transformation, where  $b$  is the translation and  $A$  is its linear part. Thus the affine maps are exactly those collineations which preserve the hyperplane of points at infinity.

**4.4.6. Determining collineations.** In order to define an affine map, it is necessary and sufficient to define an image of the affine frame. In the above description of affine transformations as special cases of projective maps, this corresponds to a suitable choice of an image of a suitable arithmetic basis of the vector space  $V$ .



In general it is not true that the image of an arithmetic basis of  $V$  determines the collineation uniquely. The basic problem is illustrated by a simple example of affine plane.

Choose four points  $A, B, C, D$  in the plane such that no three of them lie on a line. Then choose their images in the collineation as follows:

Choose arbitrarily their four images  $A', B', C', D'$  with the same property, and choose their homogeneous coordinates  $u, v, w, z, u', v', w', z'$  in  $\mathbb{R}^3$ . The vectors  $z$  and  $z'$  can be written as linear combinations

$$z = c_1u + c_2v + c_3w, \quad z' = c'_1u' + c'_2v' + c'_3w',$$

where all six coefficients must be nonzero, otherwise there exist three points not in general position.

Now choose new arithmetic representatives  $\tilde{u} = c_1u$ ,  $\tilde{v} = c_2v$  and  $\tilde{w} = c_3w$  of points  $A, B$  and  $C$  respectively. Similarly  $\tilde{u}' = c_1u'$ ,  $\tilde{v}' = c_2v'$  and  $\tilde{w}' = c_3w'$  for points  $A', B'$  and  $C'$ . This choice defines an unique linear map  $\varphi$  which maps successively

$$\varphi(\tilde{u}) = \tilde{u}', \quad \varphi(\tilde{v}) = \tilde{v}', \quad \varphi(\tilde{w}) = \tilde{w}'.$$

But then,

$$\varphi(z) = \varphi(\tilde{u} + \tilde{v} + \tilde{w}) = \tilde{u}' + \tilde{v}' + \tilde{w}' = z',$$

and so the constructed collineation maps the points which we have chosen in advance. The linear map  $\varphi$  is given uniquely by the construction, thus the collineation is given uniquely by this choice.

The argument holds also in the case when some of the chosen points are infinite (i.e. one or two). The same phenomenon can be explained even more easily by the regular collineation of a projective line. These are defined by pairwise different images of three pairwise different points.

The procedure works in an arbitrary dimension  $n$ . Then we say that  $n + 2$  points are in *general position* if no  $n + 1$  of them lie in the same hyperplane. We also call these points linearly independent, forming a *geometric basis* of projective space.

**Theorem.** *A regular collineation on  $n$ -dimensional projective space is uniquely determined by linearly independent images of  $n + 2$  linearly independent points.*

**PROOF.** The proof is exactly the same as in dimension two. We recommend writing it in detail as an exercise.  $\square$

**4.4.7. Cross-ratio.** Recall that affine maps preserve ratios of lengths of line segments on each line. Technically, we defined this ratio as for three points  $A, B$  and  $C \neq B, C = rA + sB$  as  $\lambda = (C; A, B) = -\frac{s}{r}$ .



For central projection the ratios are not preserved. Moreover, even the relative position of points on a line is not necessarily preserved. On the contrary we may determine uniquely a projective transformation by choosing arbitrarily images of three pairwise different points on a projective line. However, one can show relatively easily that the ratio of such ratios for two distinct points  $C$  is preserved:

Consider four distinct points  $A, B, C, D$  in projective space with arithmetic coordinates  $x, y, w, z$  respectively which lie on a projective line. Since these four vectors lie in the subspace generated by  $\langle x, y \rangle$ , we may write  $w$  and  $z$  as linear combinations

$$w = t_1x + s_1y, \quad z = t_2x + s_2y.$$

Define the *cross-ratio of four points*  $(A, B, C, D)$  as

$$\rho = \frac{s_1 t_2}{t_1 s_2}.$$

The definition is valid, since although the vectors  $x$  and  $y$  are determined up to a scalar multiple, these freedom is cancelled out in the definition.

Each projective transformation preserves cross-ratios: Indeed, if the transformation is given in arithmetic coordinates by a matrix  $A$ , we have images  $A \cdot w = t_1 A \cdot x + s_1 A \cdot y$ , and similarly for  $Az$ . Therefore the four images have the same cross-ratio.

We discuss the characterization of projective transformations. These are exactly those maps which preserve cross-ratios. But this is not a very practical characterization, since it contains implicitly the claim that these maps map projective lines to projective lines.

We shall not go into details here, but one can prove a much stronger statement. A map of arbitrarily small open area in affine space  $\mathbb{R}^n$  (e.g. a ball without boundary) into

the same affine space which maps lines to lines is actually a restriction of a uniquely determined globally defined collineation of the projective extension  $\mathcal{P}\mathbb{R}^{n+1}$  of the former affine space  $\mathbb{R}^n$ . As we heard above, all collineations of real projective space of finite dimensions at least two are projective transformations in our sense and thus these transformations also preserve cross-ratios.

**4.4.8. Duality.** Projective hyperplanes in an  $n$ -dimensional projective space  $\mathcal{P}(V)$  are defined as the projectivizations of  $n$ -dimensional vector subspaces in the vector space  $V$ . Hence in homogeneous coordinates, they are defined as kernels of linear forms  $\alpha \in V^*$  which in turn are determined up to a scalar multiple.



Thus in a chosen arithmetic basis, a projective hyperplane is given by a row vector  $\alpha = (\alpha_0, \dots, \alpha_n)$ . But the forms  $\alpha$  are given uniquely up to a scalar multiple. Therefore, each hyperplane in  $V$  is identified with exactly one geometric point in the projectivization of the dual space  $\mathcal{P}(V^*)$ . We call such a space the *dual projective space*, and we talk about a duality between points and hyperplanes.

On forms, the linear map defining a given collineation acts by the multiplication of row vectors from the right by the same matrix

$$\alpha = (\alpha_0, \dots, \alpha_n) \mapsto \alpha \cdot A.$$

Viewing the coordinates  $\alpha_i$  as columns, this corresponds to the action of the dual map with the matrix  $A^T$ . But the dual map maps forms in the opposite direction, from the “target space” to the “initial one”. Therefore the inverse map for the collineation of  $f$  is required in order to study the effect of regular collineations on points and their dual hyperplanes. The inverse is given by the matrix  $A^{-1}$ . Hence the matrix for the action of the corresponding collineation on forms is  $(A^T)^{-1}$ . Since the inverse matrix equals the algebraically adjoint matrix  $A_{\text{alg}}^*$ , up to the multiplication by the inverse of determinant, (see equation (1) on page 106) we can work directly with the projective transformation of the space  $\mathcal{P}(V^*)$  given by the matrix  $(A_{\text{alg}}^*)^T$  (or without transposing if we multiply row vectors from the right).

The projective point  $X$  belongs to the hyperplane  $\alpha$  if the arithmetic coordinates satisfy  $\alpha \cdot x = 0$ . It still holds after acting with an arbitrary collineation, since

$$(\alpha \cdot A^{-1}) \cdot (A \cdot x) = \alpha \cdot x = 0.$$

**4.4.9. Fixed points, centers and axes.** Consider a regular collineation  $f$  given in an arithmetic basis of projective space  $\mathcal{P}(V)$  by a matrix  $A$ .



By the *fixed point* of the collineation  $f$ , we mean a point  $P$  which is mapped to itself. That is,  $f(P) = P$ . By the *fixed hyperplane* of collineation  $f$  is meant a hyperplane  $\alpha$  which is mapped to itself. That is,  $f(\alpha) \subset \alpha$ .

Hence the arithmetic representatives of fixed points are exactly the eigenvectors of the matrix  $A$ .



In the geometry of the plane, we meet many types of collineations: reflection through a point, reflection across a line, translation, homothety etc. Perhaps we remember also some types of projections, e.g. the projection of a plane in  $\mathbb{R}^3$  to another from a center  $S \in \mathbb{R}^3$ .

Note also that there appear fixed lines next to fixed points in all cases of such affine maps. For example, the reflection through a point preserves also all lines passing through this point. In the case of a translation the infinite points behave similarly.

Now we discuss this phenomenon in an arbitrary dimension. First, we define a classical notion related to the incidence of points and hyperplanes.

A set of hyperplanes passing through a point  $P \in \mathcal{P}(V)$  is a set of all hyperplanes which contain the point  $P$ . For each point  $P$  the corresponding set of hyperplanes itself is a hyperplane in the dual space  $\mathcal{P}(V^*)$ . It is given by one homogeneous linear equation in arithmetic coordinates.

For a collineation  $f : \mathcal{P}(V) \rightarrow \mathcal{P}(V)$ , a point  $S \in \mathcal{P}(V)$ , is called the *center of collineation*  $f$ , if all hyperplanes in the set determined by  $S$  are fixed hyperplanes. A hyperplane  $\alpha$  is called the *axis of collineation*  $f$  if all its points are fixed points.

It follows that the axis of a collineation is the center of the dual collineation, while the set of hyperplanes defining the center of collineation is the axis of the dual collineation.

Since the matrices of a collineation on the former and the dual space differ only by the transposition, their eigenvalues coincide (the eigenvectors are column vectors, respectively row vectors corresponding to the same eigenvalues). For example in the projective plane (and for the same reason in each real projective space of even dimension) each collineation has at least one fixed point, since the characteristic polynomials of corresponding linear maps are of odd degree. Hence they have at least one real root.

Instead of discussing a general theory, we illustrate its usefulness in several results for projective planes.

**Proposition.** *A projective transformation of the projective plane other than the identity has either exactly one center and exactly one axis, or it has neither a center nor an axis.*

**PROOF.** Consider a collineation  $f$  on  $\mathcal{P}\mathbb{R}^3$  and assume that it has two distinct centers  $A$  and  $B$ . Denote by  $\ell$  the line given by these two centers, and choose a point  $X$  in the projective plane outside of  $\ell$ . If  $p$  and  $q$  are the lines passing through pairs of points  $(A, X)$  respectively  $(B, X)$ , then  $f(p) = p$  and  $f(q) = q$ . In particular,  $X$  is fixed. But then all points of the plane outside of  $\ell$  are fixed. Hence each line different from  $\ell$  has all points out of  $\ell$  fixed and thus also its intersection with  $\ell$  is fixed. It follows that  $f$  is the identity mapping. So it is proved that every projective transformation other than the identity has at most one center. The same argument for the dual projective plane proves that there is at most one axis.

If  $f$  has a center  $A$ , then all lines passing through  $A$  are fixed. They correspond therefore to a two-dimensional subspace of a row eigenvectors of the matrix corresponding to the transformation  $f$ . Therefore, there exists a two-dimensional subspace of column eigenvectors for the same eigenvalue. This represents exactly the line of fixed points, hence it represents the axis. The same consideration in the reversed order proves the opposite statement – if a projective transformation of plane has an axis, then it has also a center.  $\square$

picture missing!

For practical problems it is useful to work with complex projective extensions also in the case of a real plane. Then the geometric behaviour can be easily read off the potential existence of real or imaginary centers and axes.

**4.4.10. Pappus Theorem.** The following result known as Pappus theorem is a classic result of projective geometry.

**Proposition.** *Let two triples of distinct consecutive collinear points  $\{A, B, C\}$  and  $\{A', B', C'\}$  lie on two lines that meet at the point  $T$ , which is closest to  $A$  and  $A'$ , respectively. Define points  $Q, R$  and  $S$  as*

$$Q = [AB'] \cap [BA'], \quad R = [AC'] \cap [CA'], \quad S = [BC'] \cap [CB'].$$

*Then  $\{Q, R, S\}$  are also collinear.*

picture missing!

PROOF. Without loss of generality, consider the plane, passing through  $\{T, A, B, C, A', B', C'\}$  as a 2-dimensional plane in  $\mathcal{P}^2$  defined by  $z = 1$  in the homogeneous coordinates  $(x : y : z)$ .

The points  $\{T, A, B, C, A', B', C'\}$  may be considered as objects in  $\mathcal{P}^2$ , representing lines through the origin in  $\mathbb{R}^3$  with directional vectors  $\{t, a, b, c, a', b', c'\}$ , respectively. These can be chosen up to a real non-zero factor. The condition  $\{z = 1\}$  uniquely identifies those points in  $\mathbb{R}^3$  regardless of the choice of  $\{t, a, b, c, a', b', c'\}$ . Since  $\{T, A, B, C\}$  are collinear points, (they lie in the same 2-dimensional linear subspace of  $\mathbb{R}^3$ ), we may assume that this plane is generated by  $t$  and  $a$ . Choose

$$b = t + a, \quad c = \lambda t + a,$$

and analogously, for  $\{T, A', B', C'\}$

$$b' = t + a', \quad c' = \lambda' t + a'$$

for some real constants  $\lambda$  and  $\lambda'$ . It is only necessary to show that the vectors  $q, r, s$ , representing  $Q, R, S$  in  $\mathcal{P}^2$  generate a 2-dimensional subspace in  $\mathbb{R}^3$ .

Since

$$(t + a) + a' = a + (t + a'),$$

$q = t + a + a'$  represents  $Q$ . Since

$$\lambda\lambda't + \lambda'a + \lambda a' = \lambda(\lambda't + a') + \lambda'a = \lambda'(\lambda t + a) + \lambda a',$$

$r = \lambda\lambda't + \lambda'a + \lambda a'$  represents  $R$ . Finally,

$$\begin{aligned} s &= q - r = t + a + a' - \lambda\lambda't - \lambda'a - \lambda a' \\ &= (1 - \lambda')(t + a) + (1 - \lambda)(\lambda't + a') \\ &= (1 - \lambda)(t + a') + (1 - \lambda')(\lambda t + a) \end{aligned}$$

represents the point  $S$ . Thus, the points  $\{Q, R, S\}$  lie in the 2-dimensional subspace generated by vectors  $q$  and  $r$ . Since  $Q, R, S$  also belong to the plane  $\{z = 1\}$ , these points are collinear.  $\square$

#### 4.4.11. Projective classification of quadrics.

To end this section, we return to conics and quadrics. A quadric  $Q$  in  $n$ -dimensional affine space  $\mathbb{R}^n$  is defined by a general quadratic equation 4.2.6(1) on the page 342. By viewing the affine space  $\mathbb{R}^n$  as affine coordinates in projective space  $\mathcal{P}\mathbb{R}^{n+1}$  we may wish to describe the set  $Q$  by homogeneous coordinates in projective space. The formula in these coordinates should contain only the terms of second order since only a homogeneous formula is independent of the choice of the multiple of homogeneous coordinates  $(x_0, x_1, \dots, x_n)$  of a point. Hence we search for a homogeneous formula whose restriction to affine coordinates, (that is, substitution  $x_0 = 1$ ), gives the original formula 4.2.6(1).

But this is especially easy. Simply add the right number of  $x_0$  to all terms – nothing to the quadratic terms, one to the linear terms and  $x_0^2$  to the constant term in the original affine equation for  $Q$ . We obtain a well defined quadratic form  $f = \sum_{i,j=0}^n a_{ij}x_i x_j$  on the vector space  $\mathbb{R}^{n+1}$  whose zero set defines correctly the *projective quadric*  $\bar{Q}$ .

The intersection of a “cone”  $\bar{Q} \subset \mathbb{R}^{n+1}$  of the zero set of this form with the affine plane  $x_0 = 1$  is the original quadric  $Q$  whose points are called the proper points of the quadric. The other points  $\bar{Q} \setminus Q$  in the projective extension are the infinite points.

The classification of real or complex projective quadrics, up to projective transformations, is a problem already considered. It is all about finding the canonical polar basis, see paragraph 4.3.2. From this classification, given by the signature of the form in the real case and by the rank only in the complex case, we can deduce also the classification of the affine quadrics. We show the essential part of the procedure in the case of conics in the affine and the projective plane.

The projective classification gives the following possibilities, described by homogeneous coordinates  $(x : y : z)$  in the projective plane  $\mathcal{P}\mathbb{R}^3$ :

- imaginary regular conic given by  $x^2 + y^2 + z^2 = 0$
- real regular conic given by  $x^2 + y^2 - z^2 = 0$
- pair of imaginary lines given by  $x^2 + y^2 = 0$
- pair of real lines given by  $x^2 - y^2 = 0$
- one double line  $x^2 = 0$ .

We consider this classification as real, that is, the classification of quadratic forms is given not only by its rank but also by its signature. Nevertheless, the points of a quadric are considered also in the complex extension. In this way we should understand the stated names. For example the imaginary conic does not have any real points.

**4.4.12. Affine classification of quadrics.** For an affine classification we must restrict the projective transformations to those which preserve the projective subspace of infinite points. This can be seen also by the converse procedure — for a fixed projective type of conic  $Q$ , that is, its cone  $\tilde{Q} \subset \mathbb{R}^3$ , we choose different affine planes  $\alpha \subset \mathbb{R}^3$  which do not pass through the origin. We observe the changes to the set of points  $\tilde{Q} \cap \alpha$ , which are proper points of  $Q$  in affine coordinates, as realized by the plane  $\alpha$ .

Hence in the case of a regular conic there is a real cone  $\tilde{Q}$  given by the equation  $z^2 = x^2 + y^2$ . As planes  $\alpha$  we may for instance choose the tangent planes to the unit sphere. If we begin with the plane  $z = 1$ , the intersection consists only of finite points forming a unit circle  $Q$ . By a gradual change of the slope of  $\alpha$  we obtain a more and more stretched ellipse until we get such a slope that  $\alpha$  is parallel with one of lines of the cone. At that moment there appears one (double) infinite point of the conic whose finite points still form one connected component, and so we have a parabola. Continuing to change the slope gives rise to two infinite points. The set of finite points is no longer connected, and so we obtain the last regular quadric in the affine classification, a hyperbola.

We can take advice from the introduced method which enables us to continue the classification in higher dimensions. In particular, we notice that the intersection of the conic with the projective line of infinite points is always a quadric in dimension one less. It is either the empty set or a double point

or two points as types of quadrics on a projective line. Next we found an affine transformation transforming one of possible realizations of a fixed projective type to another one, only if the corresponding quadrics in the subspace of infinite points were projectively equivalent. In this way, it is possible to continue the classification of quadrics in dimension three and above.

**D. Further exercise on this chapter**

**4.D.1.** [TO BE MOVED TO AN APPROPRIATE SPOT IN CHAPTER 8 OR 9] Show that all twice differentiable mappings  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  which preserve distances are affine mappings, and thus Euclidean (also the so called rigid motions).

**Solution.** Suppose  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  keeps distances, i.e.,  $\|F(x + tv) - F(x)\| = \|tv\|$ , where  $t \in \mathbb{R}$ ,  $v \in \mathbb{R}^n$ . Then the Taylor expansion says

$$t\|v\| = \|F(x+tv) - F(x)\| = \|DF(x)(tv) + \frac{1}{2}D^2F(x+s_tv)(tv, tv)\| = t\|DF(x)(v) + \frac{t}{2}D^2F(x+s_tv)(v, v)\|,$$

where  $0 \leq s_t \leq t$ . Now, the limit for  $t \rightarrow 0$  leads to  $\|v\| = \|DF(x)(v)\|$ , i.e., the expected fact that the differential of  $F$  must be an orthogonal linear map. Consequently, writing  $F_{x_i}$  for the partial derivatives, its columns are orthogonal:

$$(1) \quad F_{x_i} \cdot F_{x_j} = \delta_{ij}.$$

In particular,  $F$  is invertible on a neighborhood of  $x$  and we may look at  $G(y) = DF(x)^{-1} \circ F(y)$  instead of  $F$ . Now, the differential of  $G$  is the identity matrix at the point  $x$  and  $G$  keeps distances again. Thus, differentiating the equation (1) for  $G$  we arrive at

$$G_{x_i x_k}(x) \cdot G_{x_j}(x) + G_{x_i}(x) \cdot G_{x_j x_k}(x) = 0$$

and taking into account  $G_{x_i}^j(x) = \delta_i^j$ , the latter equation reduces to  $\alpha_{jik} = -\alpha_{ijk}$ , where we write  $\alpha_{ijk} = G_{x_j x_k}^i(x)$  for the second partial derivatives of the relevant component function. Clearly, we also know  $\alpha_{ijk} = \alpha_{ikj}$  since the second derivatives are commutative. Thus,

$$\alpha_{ijk} = -\alpha_{jik} = -\alpha_{jki} = \alpha_{kji} = \alpha_{kij} = -\alpha_{ikj} = -\alpha_{ij k}$$

and thus all the quantities  $\alpha_{ijk}$  have to vanish. This means all the second partial derivatives of  $G$  in the point  $x$  vanish and by the definition, this also implies that all the second order partial derivatives of  $F(y) = DF(x) \circ G(y)$  in the point  $x$  vanish (write down the formulae explicitly in case of any doubts!). The point  $x$  was arbitrary and thus we have verified that  $F$  is affine by the Taylor expansion. □

**4.D.2.** Find a parametric equation for the intersection of the following planes in  $\mathbb{R}^3$ :

$$\sigma : 2x + 3y - z + 1 = 0 \quad \text{a} \quad \rho : x - 2y + 5 = 0.$$



**4.D.3.** Find a common perpendicular for the skew lines

$$p : [1, 1, 1] + t(2, 1, 0), \quad q : [2, 2, 0] + t(1, 1, 1).$$



**4.D.4.** Jarda is standing in  $[-1, 1, 0]$  and has a stick of length 4. Can he simultaneously touch the lines  $p$  and  $q$ , where

$$\begin{aligned} p & : [0, -1, 0] + t(1, 2, 1), \\ q & : [3, 4, 8] + s(2, 1, 3)? \end{aligned}$$

(The stick must pass through  $[-1, 1, 0]$ .)



**4.D.5.** A cube  $ABCDEFGH$  is given. The point  $T$  lies on the edge  $BF$ , with  $|BT| = \frac{1}{4}|BF|$ . Compute the cosine of the angle between  $ATC$  and  $BDE$ . □



**4.D.6.** A cube  $ABCDEFGH$  is given. The point  $T$  lies on the edge  $AE$ , with  $|AT| = \frac{1}{4}|AE|$ .  $S$  is the midpoint of  $AD$ . Compute the cosine of the angle between  $BDT$  and  $SCH$ . □



**4.D.7.** A cube  $ABCDEFGH$  is given. The point  $T$  lies on the edge  $BF$ ,  $|BT| = \frac{1}{3}|BF|$ . Compute the cosine of the angle between  $ATC$  and  $BDE$ . □



**4.D.8.** What are the lengths of semi-axes, when the sum of their lengths equals the distance between foci both equal 1.

**Solution.** It is given that  $a + b = 1$  and  $2ae = 1$ . Also  $b^2 = a^2(1 - e^2)$ . Eliminating  $e$  gives  $b^2 = a^2 - (1/4)$ . So  $1/4 = a^2 - b^2 = (a - b)(a + b) = a - b$ . So  $a = 5/8$  and  $b = 3/8$ . □

**Solution.** (Alternative.) Solve the system

$$\begin{aligned} a + b & = 1 \\ 2e & = 2\sqrt{a^2 - b^2} = 1 \end{aligned}$$

and find solution  $a = \frac{5}{8}$ ,  $b = \frac{3}{8}$ . □

**4.D.9.** For what slopes  $k$  are the lines passing through  $[-4, 2]$  secant and tangent lines of the ellipse defined by

$$\frac{x^2}{9} + \frac{y^2}{4} = 1$$

**Solution.** The direction vector of the line is  $(1, k)$  and its parametric equations then are  $x = -4 + t$ ,  $y = 2 + kt$ . The intersection with the ellipse satisfies

$$\frac{(-4 + t)^2}{9} + \frac{(2 + kt)^2}{4} = 1$$

This quadratic equation has discriminant equal to

$$D = -\frac{k}{9}(7k + 16).$$

This implies that for  $k \in (-\frac{16}{7}, 0)$  there are two solutions, and the line is a secant. For  $k = -\frac{16}{7}$  and  $k = 0$  there is only one solution and the line is a tangent to the ellipse.  $\square$

**4.D.10.** Find all lines tangent to the ellipse  $3x^2 + 7y^2 = 30$ , so that the distance from the centre of the ellipse to the tangent is 3.

**Solution.** All lines at distance 3 from the origin are tangents to the circle centre at  $[0, 0]$  and radius 3. They all have an equation  $x \cos \theta + y \sin \theta = 3$  for some  $\theta$ . This line meets the standard ellipse  $x^2/a^2 + y^2/b^2 = 1$  where

$$\frac{x^2}{a^2} + \frac{(3 - x \cos \theta)^2}{b^2 \sin^2 \theta} = 1$$

or

$$x^2(a^2 \cos^2 \theta + b^2 \sin^2 \theta) - 6a^2 x \cos \theta - a^2(b^2 \sin^2 \theta - 9) = 0$$

It is a tangent line if the above equation has a double root for  $x$ . Thus it is required that

$$36a^4 \cos^2 \theta = 4(a^2 \cos^2 \theta + b^2 \sin^2 \theta)(9 - b^2 \sin^2 \theta).$$

This simplifies to requiring that

$$a^2 \cos^2 \theta + b^2 \sin^2 \theta = 9.$$

This implies

$$\cos^2 \theta = \frac{(9 - b^2)}{(a^2 - b^2)}$$

$$\sin^2 \theta = \frac{(a^2 - 9)}{(a^2 - b^2)}$$

For the given problem  $a^2 = 10$  and  $b^2 = 30/7$ . The solution is  $x\sqrt{33} + y\sqrt{7} = 3\sqrt{40}$ .  $\square$

**Solution.** (Alternative.) The tangent line is  $(y - b \sin \theta) = -\frac{b \cos \theta}{a \sin \theta}(x - a \cos \theta)$  with  $a^2 = 10$  and  $b^2 = 30/7$ . The distance to the origin, 3, implies  $3 = (a/\cos \theta) \sin \varphi$  where

$$\tan \varphi = \frac{b \cos \theta}{a \sin \theta} \quad 3 \cos \theta = a \sin \varphi \quad \text{where} \quad a \sin \theta \tan \varphi = b \cos \theta \quad 3 \sin \theta = b \cos \varphi$$

$$9/a^2 \cos^2 \theta + 9/b^2 \sin^2 \theta = 1. \quad 9 \cos^2 \theta + 21 \sin^2 \theta = 10. \quad 12 \sin^2 \theta = 1$$

$$(y - \sqrt{35/2}) = -(x - \sqrt{55}/\sqrt{6})[\sqrt{33/7}] \quad \square$$

**Solution.** (Alternative.) The centre of the ellipse is at the origin. The distance  $d$  between the line  $ax + by + c = 0$  and the origin is  $d = \frac{|c|}{\sqrt{a^2 + b^2}}$ . The tangent then satisfies  $a^2 + b^2 = \frac{c^2}{9}$ . The equation of the tangent passing through the point  $[x_T, y_T]$  is  $3xx_T + 7yy_T - 30 = 0$ . For coordinates of the point of tangency,

$$\begin{aligned} (3x_T)^2 + (7y_T)^2 &= 100 \\ 3x_T^2 + 7y_T^2 &= 30 \end{aligned}$$

Its solution is  $x_T = \pm\sqrt{\frac{55}{6}}$ ,  $y_T = \pm\sqrt{\frac{5}{14}}$ . Considering the symmetry of ellipse, there are four solutions

$$\pm 3\sqrt{\frac{55}{6}}x \pm 7\sqrt{\frac{5}{14}}y - 30 = 0. \quad \square$$

**4.D.11.** A hyperbola  $x^2 - y^2 = 2$  is given. Find an equation of a hyperbola having the same foci and passing through point  $[-2, 3]$ .

**Solution.** The given hyperbola has  $a^2 = b^2 = 2$ , so  $a^2e^2 = a^2 + b^2 = 4$ , and the foci are at  $(\pm ae, 0) = (\pm 2, 0)$ . So the desired hyperbola has equation

$$\sqrt{(x-2)^2 + y^2} - \sqrt{(x+2)^2 + y^2} = k,$$

for some constant  $k$ . Since the hyperbola passes through  $[-2, 3]$ ,  $k = 2$ . Squaring gives

$$\sqrt{(x-2)^2 + y^2} = [\sqrt{(x+2)^2 + y^2} + 2],$$

$$(x-2)^2 + y^2 = (x+2)^2 + y^2 + 4\sqrt{(x+2)^2 + y^2} + 4$$

$(-2x-1)^2 = (x+2)^2 + y^2$  or  $3x^2 = y^2 + 3$  which is the required hyperbola. □

**Solution.** (Alternative.) The equation of the desired hyperbola is  $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ , with its eccentricity  $e$  satisfying  $a^2e^2 = a^2 + b^2 = 4$ , since the foci are at  $(\pm ae, 0) = (\pm 2, 0)$ . The point  $[-2, 3]$  lies on the hyperbola, so  $\frac{4}{a^2} - \frac{9}{b^2} = 1$ . It follows that  $a^2 = 1$ ,  $b^2 = 3$ . The desired hyperbola is  $x^2 - \frac{y^2}{3} = 1$ . □

**4.D.12.** Determine the equations of the tangent lines to the hyperbola  $4x^2 - 9y^2 = 1$ , which are perpendicular to line  $x - 2y + 7 = 0$ .

**Solution.** All lines perpendicular to the given line have an equation  $2x + y + c = 0$  for some  $c$ . So the line has an intersection with a double root with the given hyperbola. So the equation  $4x^2 - 9(-2x - c)^2 = 1$  has a double root. Hence  $(36c)^2 - 4 \cdot 32 \cdot (9c^2 + 1) = 0$ , and  $c = \pm \frac{2\sqrt{2}}{3}$ . □

**4.D.13.** Determine the tangent to the ellipse  $\frac{x^2}{16} + \frac{y^2}{9} = 1$  which is parallel with line  $x + y - 7 = 0$ .

**Solution.** The lines parallel with the given line intersect this line in a point at infinity  $(1 : -1 : 0)$ . Construct tangents to given ellipse passing through this point. The point of tangency  $T = (t_1 : t_2 : t_3)$  lies on its polar and therefore satisfies  $\frac{t_1}{16} - \frac{t_2}{9} = 0$ , so  $t_2 = \frac{9}{16}t_1$ . Substituting into the ellipse equation, we get  $t_1 = \pm \frac{16}{5}$ . The touching points of the desired tangents are  $[\frac{16}{5}, \frac{9}{5}]$  and  $[-\frac{16}{5}, -\frac{9}{5}]$ . The tangents are polars of those points. They have equations  $x + y = 5$  and  $x + y = -5$ . □

**Solution.** (Alternative.) The given line has slope  $-1$ . The tangent line at  $(4 \cos \theta, 3 \sin \theta)$  has slope  $-\frac{3 \cos \theta}{4 \sin \theta}$ , so it is required that  $\tan \theta = \frac{3}{4}$ . The tangent line has equation  $(y - 3 \sin \theta) = (-1)(x - 4 \cos \theta)$  where either  $\sin \theta = 3/5$  and  $\cos \theta = 4/5$  or  $\sin \theta = -3/5$  and  $\cos \theta = -4/5$ . The two solutions are  $x + y = \pm 5$ . □

**4.D.14.** Determine the points at infinity and the asymptotes of the conic section

$$2x^2 + 4xy + 2y^2 - y + 1 = 0$$

**Solution.** The equation for the points at infinity of  $2x^2 + 4xy + 2y^2 = 0$  or rather  $2(x + y)^2 = 0$  has a solution  $x = -y$ . The only point at infinity therefore is  $(1 : -1 : 0)$ , so the conic section is a parabola. The asymptote is a polar of this point, specifically the line at infinity  $z = 0$ . □

**4.D.15.** Prove that the product of the distances between an arbitrary point on a hyperbola and both of its asymptotes is constant. Find its value.

**Solution.** Denote the point lying on the hyperbola by  $P$ . The asymptote equation of the hyperbola in canonical form is  $bx \pm ay = 0$ . Their normals are  $(b, \pm a)$  and from here we determine the projections  $P_1, P_2$  of point  $P$  to asymptotes. For the distance between point  $P$  and asymptotes we get  $|PP_{1,2}| = \frac{|aq \pm bp|}{\sqrt{a^2 + b^2}}$ . The product is therefore equal to  $\frac{a^2q^2 - b^2p^2}{a^2 + b^2} = \frac{a^2b^2}{a^2 + b^2}$ , because  $P$  lies on hyperbola. □

**4.D.16.** Compute the angle between the asymptotes of the hyperbola  $3x^2 - y^2 = 3$ .

**Solution.** For the cosine of the angle between the asymptotes of the hyperbola in canonical form,  $\cos \alpha = \frac{b^2 - a^2}{b^2 + a^2}$ . In this case the angle is 60 degrees. □



**4.D.17.** Locate the centers of the conic sections

- (a)  $9x^2 + 6xy - 2y - 2 = 0$   
 (b)  $x^2 + 2xy + y^2 + 2x + y + 2 = 0$   
 (c)  $x^2 - 4xy + 4y^2 + 2x - 4y - 3 = 0$   
 (d)  $\frac{(x-\alpha)^2}{a^2} + \frac{(y-\beta)^2}{b^2} = 1$

**Solution.** (a) The system  $\bar{A}s + a = 0$  for computing centers is

$$\begin{aligned} 9s_1 + 3s_2 &= 0 \\ 3s_1 - 2 &= 0 \end{aligned}$$

Solve it to obtain the center at  $[\frac{2}{3}, -2]$ .

(b) In this case,

$$\begin{aligned} s_1 + s_2 + 1 &= 0 \\ s_1 + s_2 + \frac{1}{2} &= 0. \end{aligned}$$

Therefore there is no proper center (the conic section is a parabola). Moving to homogeneous coordinates we can obtain the center at infinity  $(1 : -1 : 0)$ .

(c) The coordinates of the center in this case satisfy

$$\begin{aligned} s_1 - 2s_2 + 1 &= 0 \\ -2s_1 + 4s_2 - 2 &= 0. \end{aligned}$$

The solution is the line of centers. This is so because the conic section is degenerate: it is a pair of parallel lines.

(d) The center is at  $(\alpha, \beta)$ . The coordinates of the center therefore give the translation of the coordinate system to the frame in which the ellipse has its basic form. □

**4.D.18.** Find the equations of the axes of the conic section  $6xy + 8y^2 + 4y + 2x - 13 = 0$ .

**Solution.** The major and minor axes of the conic section are in the direction of the eigenvectors of matrix  $\begin{pmatrix} 0 & 3 \\ 3 & 8 \end{pmatrix}$ . The characteristic equation has the form  $\lambda^2 - 8\lambda - 9 = 0$ . The eigenvalues are therefore  $\lambda_1 = -1, \lambda_2 = 9$ . The corresponding eigenvectors are then  $(3, -1)$  and  $(1, -3)$ . The axes are the polars of points at infinity defined by those directions. For  $(3, -1)$ , the axis equation is  $-3x + y + 1 = 0$ . For  $(1, -3)$  it is  $-9x - 21y - 5 = 0$ . □

**4.D.19.** Determine the equations of the axes of the conic section  $4x^2 + 4xy + y^2 + 2x + 6y + 5 = 0$ .

**Solution.** The eigenvalues of the matrix  $\begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$  are  $\lambda_1 = 0, \lambda_2 = 5$  and the corresponding eigenvectors are  $(-1, 2)$  and  $(2, 1)$ . There is one axis  $2x + y + 1 = 0$ , and the conic section is a parabola. □

**4.D.20.** The equation

$$x^2 + 3xy - y^2 + x + y + 1 = 0.$$

defines a conic section. Determine its center, axes, asymptotes and foci.

**4.D.21.** Find the equation of the tangent at  $P=[1, 1]$  to the conic section

$$4x^2 + 5y^2 - 8xy + 2y - 3 = 0$$

**Solution.** By projecting, this is a conic section defined by the quadratic form  $(x, y, z)A(x, y, z)^T$  with matrix

$$A = \begin{pmatrix} 4 & -4 & 0 \\ -4 & 5 & 1 \\ 0 & 1 & -3 \end{pmatrix}$$

Using the previous theorem, the tangent is a polar of P, which has homogenous coordinates  $(1 : 1 : 1)$ . It is given by equation  $(1, 1, 1)A(x, y, z)^T = 0$ , which in this case gives

$$2y - 2z = 0$$

Moving back to inhomogeneous coordinates, the tangent line equation is  $y = 1$ .

□

**4.D.22.** Find the coordinates of the intersection of the  $y$  axis and the conic section defined by

$$5x^2 + 2xy + y^2 - 8x = 0$$

**Solution.** The  $y$  axis, is the line  $x = 0$ . It is the polar of the point P with homogeneous coordinates  $\langle p \rangle = (p_1 : p_2 : p_3)$ . That means that the equation  $x = 0$  is equivalent to the polar equation  $F(p, v) = p^T Av = 0$ , where  $v = (x, y, z)^T$ . This is satisfied when  $Ap = (\alpha, 0, 0)^T$  for some  $\alpha \in \mathbb{R}$ . This condition gives the conic section matrix

$$A = \begin{pmatrix} 5 & 1 & -4 \\ 1 & 1 & 0 \\ -4 & 0 & 0 \end{pmatrix}$$

equation system

$$\begin{aligned} 5p_1 + p_2 - 4p_3 &= \alpha j \\ p_1 + p_2 &= 0 \\ -4p_1 &= 0 \end{aligned}$$

We can find the coordinates of P by the inverse matrix,  $p = A^{-1}(\alpha, 0, 0)^T$ , or solve the system directly by backward substitution. In this case we can easily obtain solution  $p = (0, 0, -\frac{1}{4}\alpha)$ . So the  $y$  axis touches the conic section at the origin.

□

**4.D.23.** Find a touch point of the line  $x = 2$  with the conic section from the previous exercise.

**Solution.** The line has equation  $x - 2z = 0$  in its projective extension and therefore we get the condition  $Ap = (\alpha, 0, -2\alpha)$  for the touch point P, which gives

$$\begin{aligned} 5p_1 + p_2 - 4p_3 &= \alpha \\ p_1 + p_2 &= 0 \\ -4p_1 &= -2\alpha \end{aligned}$$

Its solution is  $p = (\frac{1}{2}\alpha, -\frac{1}{2}\alpha, \frac{1}{4}\alpha)$ . These homogeneous coordinates are equivalent to  $(2, -2, 1)$  and hence the touch point has coordinates  $[2, -2]$ .

□

**4.D.24.** Find equations of the tangents passing through  $P = [3, 4]$  to the conic defined by

$$2x^2 - 4xy + y^2 - 2x + 6y - 3 = 0.$$

**Solution.** Suppose that the point of tangency  $T$  has homogeneous coordinates given by a multiple of the vector  $t = (t_1, t_2, t_3)$ . The condition that  $T$  lies on the conic section is  $t^T At = 0$ , which gives

$$2t_1^2 - 4t_1t_2 + t_2^2 - 2t_1t_3 + 6t_2t_3 - 3t_3^2 = 0$$

The condition that P lies on the polar of T is  $p^T At = 0$ , where  $p = (3, 4, 1)$  are the homogeneous coordinates of point P. In this case, the equation gives

$$(3, 4, 1) \begin{pmatrix} 2 & -2 & -1 \\ -2 & 1 & 3 \\ -1 & 3 & -3 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = -3t_1 + t_2 + 6t_3 = 0$$

Now we can substitute  $t_2 = 3t_1 - 6t_3$  to the previous quadratic equation. Then

$$-t_1^2 + 4t_1t_3 - 3t_3^2 = 0$$

Because the equation is not satisfied for  $t_3 = 0$ , we move to inhomogeneous coordinates  $(\frac{t_1}{t_3}, \frac{t_2}{t_3}, 1)$ , for which we get

$$-\left(\frac{t_1}{t_3}\right)^2 + 4\left(\frac{t_1}{t_3}\right) - 3 = 0 \quad \text{a} \quad \frac{t_2}{t_3} = 3\left(\frac{t_1}{t_3}\right) - 6,$$

tj.  $\frac{t_1}{t_3} = 1$  a  $\frac{t_2}{t_3} = -3$ , nebo  $\frac{t_1}{t_3} = 3$  a  $\frac{t_2}{t_3} = 3$ . So the touch points have homogeneous coordinates  $(1 : -3 : 1)$  and  $(3 : 3 : 1)$ . The tangent equations are the polars of those points  $7x - 2y - 13 = 0$  and  $x = -3$ .

□

**4.D.25.** Find an equation of the tangent passing through the origin to the circle

$$x^2 + y^2 - 10x - 4y + 25 = 0$$

**Solution.** The touch point  $(t_1 : t_2 : t_3)$  satisfies

$$(0, 0, 1) \begin{pmatrix} 1 & 0 & -5 \\ 0 & 1 & -2 \\ -5 & -2 & 25 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix} = -5t_1 - 2t_2 + 25 = 0$$

From here we eliminate  $t_2$  and substitute into circle equation, which  $(t_1 : t_2 : t_3)$  has to be satisfied as well. We obtain the quadratic equation  $29t_1^2 - 250t_1 + 525 = 0$ , with solutions  $t_1 = 5$  and  $t_1 = \frac{105}{29}$ . We compute the coordinate  $t_2$  and get touch points  $[5, 0]$  and  $[\frac{105}{29}, \frac{100}{29}]$ . The tangents are polars of those points with equations  $y = 0$  and  $20x - 21y = 0$ .  $\square$

**4.D.26.** Find tangents equations to circle  $x^2 + y^2 = 5$  which are parallel with  $2x + y + 2 = 0$ .

**Solution.** In the projective extension, these tangents intersect at the point at infinity satisfying  $2x + y + z = 0$ , so in point with homogeneous coordinates  $(1 : -2 : 0)$ . They are tangents from this point to the circle. We can use the same method as in previous exercise. The conic section matrix is diagonal with the diagonal  $(1, 1, -5)$  and therefore the touch point  $(t_1 : t_2 : t_3)$  of the tangents satisfies  $t_1 - 2t_2 = 0$ . Substitute into the circle equation to get  $5t_2^2 = 5$ . Since  $t_2 = \pm 1$ , the touch points are  $[2, 1]$  and  $[-2, -1]$ .  $\square$

**Solution.** Alternative. The point  $P = \sqrt{5}(\cos \theta, \sin \theta)$  lies on the circle for all  $\theta$ . The tangent line at  $P$  is  $x \cos \theta + y \sin \theta = \sqrt{5}$ . This has slope  $-(\cos \theta)/(\sin \theta)$  which is  $-2$  provided  $\tan \theta = 1/2$ . It follows that  $P$  is at either  $[2, 1]$  or  $[-2, -1]$ .  $\square$

A tangent line touching the conic section at infinity is called an *asymptote*. The number of asymptotes of a conic section equals the number of intersections between the conic section and the line at infinity. So the ellipse has no real asymptotes, the parabola has one (which is however a line at infinity) and the hyperbola has two.

**4.D.27.** Find the points at infinity and the asymptotes of the conic section defined by

$$4x^2 - 8xy + 3y^2 - 2y - 5 = 0$$

**Solution.** First, rewrite the conic section in homogeneous coordinates.

$$4x^2 - 8xy + 3y^2 - 2yz - 5z^2 = 0$$

the homogeneous coordinates  $(x : y : 0)$  satisfying this equation, which means

$$4x^2 - 8xy + 3y^2 = 0.$$

It follows that either:  $\frac{x}{y} = -\frac{1}{2}$  or  $\frac{x}{y} = -\frac{3}{2}$ . The conic section is therefore a hyperbola with points at infinity  $P = (-1 : 2 : 0)$  and  $Q = (-3 : 2 : 0)$ .

$$(-1, 2, 0) \begin{pmatrix} 4 & -4 & 0 \\ -4 & 3 & -1 \\ 0 & -1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = -12x + 10y - 2 = 0$$

and

$$(-3, 2, 0) \begin{pmatrix} 4 & -4 & 0 \\ -4 & 3 & -1 \\ 0 & -1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = -20x + 18y - 2 = 0$$

$\square$

There are further exercises on conic sections on the page 361.

**4.D.28. Harmonic cross-ratio.** If the cross-ratio of four points lying on the line equals  $-1$ , we talk about a *harmonic quadruple*. Let  $ABCD$  be a quadrilateral. Denote by  $K$  the intersection of the lines  $AB$  and  $CD$ , by  $M$  the intersection of the lines  $AD$  and  $BC$ . Further let  $L, N$  be the intersection of  $KM$  and  $AC, BD$  respectively. Show that the points  $K, L, M, N$  are a harmonic quadruple.  $\circ$

**Exercise solution**

## Establishing the ZOO

which functions do we need for our models?  
– a thorough menagerie



### A. Polynomial interpolation

Our first “route” in the zoo of functions is devoted to a beautiful topic, concerning the elementary notion of polynomials. Usually, given a polynomial one wants to find its roots, and in the first chapter we learned to evaluate polynomials via the Horner’s scheme. Of course, evaluation is an essential part of any root-finding method and we know that the fundamental theorem of algebra classifies all roots of a polynomial over  $\mathbb{C}$  (see also the paragraph 12.2.8).

Below we adopt a slightly different approach, and focus on applications related to the notion of polynomial interpolation of real or complex functions. Roughly speaking this means that we will use polynomials of certain degree to approximate such functions. This includes the so called “Lagrange interpolation”, but we will also explain how the theory of derivatives involves, in terms of “Hermite’s interpolation problem” and “cubic splines”. We mention that in this part the use of elementary functions is stressed, although they are systematically analyzed later in this chapter.

**5.A.1. Using Sage for plotting graphs.** Computer algebra programs, as Mathematica, Matlab, Maple, Sage, and other, all include simple syntaxes available for drawing the graph  $G_f = \{(x, f(x)) : x \in I\}$  of a real polynomial  $f(x) = a_n x^n + \dots + a_1 x + a_0$  of some fixed degree  $n \in \mathbb{N}$ , or

In this chapter, we start using tools allowing us to model dependencies which are neither linear, nor discrete. Such models are often needed when dealing with time dependent systems. We try to describe them not only at discrete moments of time, but “continuously”. Sometimes this is advantageous, for instance in physical models of classical mechanics and engineering. It might also be appropriate and computationally effective to employ an approximation of discrete models in economics, chemistry, or biology. In particular such ideas may be appropriate in relation to stochastic models, as we shall see in Chapter 10.

The key concept is that of a function, also called a “signal” in practical applications. The larger the class of functions used, the more difficult is the development of effective tools. On the other hand, if there are only a few simple types of functions available, it may be that some real situations cannot be modelled at all.

The objective of the following two chapters is thus to introduce explicitly the most elementary functions of real variables. It is also to describe implicitly many more functions, and to build the standard tools to use them. This is the differential and integral calculus of one variable. While the focus has been mainly on the part of mathematics called *algebra*, the emphasis will now be on *mathematical analysis*. The link between the two is provided by a “geometric approach”. If possible, this means building concepts and intuition independently of any choice of coordinates. Often this leads to a discrete (finite) description of the objects of interest. This is immediate when working with polynomials now.

### 1. Polynomial interpolation

In the previous chapters, we often worked with sequences of real or complex numbers, i.e. with scalar functions  $\mathbb{N} \rightarrow \mathbb{K}$  or  $\mathbb{Z} \rightarrow \mathbb{K}$ , where  $\mathbb{K}$  is a given set of numbers. We also worked with sequences of vectors over real or complex numbers.

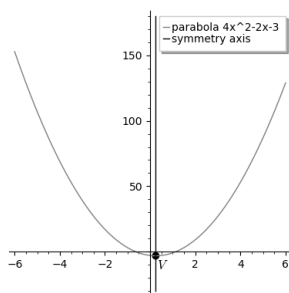
Recall the discussion from paragraph 1.1.6, about dealing with scalar functions. This discussion is adequate to work with functions  $\mathbb{R} \rightarrow \mathbb{R}$  (*real-valued functions of one real variable*), or  $\mathbb{R} \rightarrow \mathbb{C}$  (*complex-valued functions of one real variable*), or sometimes more generally the *vector-valued functions of one real variable*  $\mathbb{R} \rightarrow V$ . The results can usually be

more general of some real function  $f : I \rightarrow \mathbb{R}$ , where  $I$  is a suitable interval in the real line. Next we will use Sage, and its command `plot(f(x), x, a, b)`, a situation that we encountered already in Chapter 1.

It worths to mention that Sage computes the plot of a given function  $f$  by evaluating the function at a large number of random points lying in the chosen interval  $[a, b]$ . Since in general Sage tries to plot the whole graph of  $f$ , it is often useful to restrict the “ $y$ -range” between certain values (“compatible” with the range of  $y = f(x)$ , say  $c \leq y = f(x) \leq d$ ).<sup>1</sup> This can be done by including inside the command `plot` the options `ymin = c` and `ymax = d`, respectively. The restriction of the  $y$ -range is also useful for treating further specialities, as for example for the graphs of functions with (vertical) asymptotes and other situations. We will discuss such details later.

**5.A.2. Low degree polynomials.** Describe geometrically real polynomials of degree zero, one, two, and three, respectively. Use Sage to plot the corresponding graphs for the most “fancy cases” (i.e., the non-linear cases).

**Solution.** Let us discuss the cases of degree  $n = 2, 3$  only (whose graphs are more interesting than straight lines, being the graphs for the cases  $n = 0, 1$ ). So, recall that a second degree polynomial has the form  $f(x) = a_2x^2 + a_1x + a_0 = ax^2 + bx + c$ , with  $a_2 = a \neq 0$ , and  $x \in \mathbb{R}$ . Its graph is a parabola, symmetric with respect to the vertical line  $x = -\frac{b}{2a}$  and oriented upwards or downwards, depending on the sign of  $a$ . Below we present the graph of the parabola  $f(x) = 4x^2 - 2x - 3$ , constructed in Sage. This indicates also its axis of symmetry and its vertex  $V = [-\frac{b}{2a}, f(-\frac{b}{2a})] = [1/4, -13/4]$ .



Recall from 1.G.5 that a method to sketch the graph of  $f(x) = a_0 + a_1x + a_2x^2$  is based on the substitution method and in particular relies on a combination of the commands `plot` and `subs`. This allows to substitute certain values of the parameters  $a_0, a_1, a_2$  and quickly sketch the corresponding  $C_f$  (this can be applied to plot the graph of any polynomial).

```
a=SR.var("a", 3); P=a[0]+a[1]*x+a[2]*x^2
q=plot(P.subs(a[0]==-3,a[1]==-2,a[2]==4),
x,-6,6); q.show()
```

As a side remark related to the use of `ymin` and `ymax` mentioned in 5.A.1, observe that one may include them also in

<sup>1</sup>This is extremely useful when for example we need to sketch the graph of a rational function.

extended to cases concerning vector values over the considered scalars, rather than just real and complex numbers.

We begin with some easily computable functions.

**5.1.1. Polynomials.** We can add and multiply scalars.



These operations satisfy a number of properties which we listed in the paragraphs 1.1.1 and 1.1.5. If we admit any finite number of these operations, leaving one of the variables as an unknown and fixing the other scalars, we obtain the polynomial functions. We consider the scalars  $\mathbb{K} = \mathbb{R}, \mathbb{C}$ , or  $\mathbb{Q}$ .

POLYNOMIALS

A polynomial over a ring of scalars  $\mathbb{K}$  is a mapping  $f : \mathbb{K} \rightarrow \mathbb{K}$  given by the expression

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0,$$

where  $a_i, i = 0, \dots, n$ , are fixed scalars. Multiplication is indicated by juxtaposition of symbols, and “+” denotes addition. If  $a_n \neq 0$ , the polynomial  $f$  is said to have degree  $n$ . The degree of the zero polynomial is undefined. The scalars  $a_i$  are called the *coefficients of the polynomial*  $f$ .

The polynomials of degree zero are exactly the non-zero constant mappings. In algebra, polynomials are more often defined as formal expressions of the aforementioned form of  $f(x)$ , i. e. a polynomial is defined to be a sequence  $a_0, a_1, \dots$  of coefficients such that only finitely many of them are non-zero. However, we will show shortly that these approaches are equivalent for our choices of scalars.

It is easy to verify that the polynomials over a given ring of scalars form a ring. Multiplication and addition of polynomials are given by the operations in the original ring  $\mathbb{K}$  by the values of the polynomials. Hence,

$$(f \cdot g)(x) = f(x) \cdot g(x), \quad (f + g)(x) = f(x) + g(x),$$

where the operations on the left-hand side are interpreted in the ring of polynomials whereas the operations on the right-hand side are of the ring of scalars (see the second part of Chapter 12 for detailed algebraic treatment).

**5.1.2. Division of polynomials.** As already mentioned, the scalar fields used are  $\mathbb{Q}, \mathbb{R}$ , or  $\mathbb{C}$ . In all of these fields, the following holds:

EUCLIDEAN DIVISION OF POLYNOMIALS

**Proposition.** For any two polynomials  $f$  of degree  $n$  and  $g$  of degree  $m$ , there is exactly one pair of polynomials  $q, r$  such that  $f = q \cdot g + r$ , where either  $r = 0$ , or the degree of  $r$  is less than  $m$ .



**PROOF.** This is a special simple case of the much more general algebraic result in 12.3.6. Write  $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$  for the polynomial of degree  $n$ , and  $g(x) = b_mx^m + b_{m-1}x^{m-1} + \dots + b_1x + b_0$ , with  $a_n \neq 0$  and  $b_m \neq 0$ .

the command `show` ending the previous block, to restrict the values of  $f$ . For  $n = 3$ , the graph of a general cubic polynomial  $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$ , with  $a_3 \neq 0$  is a curve running from  $-\infty$  to  $\infty$ , if  $a_3 > 0$ , or the opposite way from  $\infty$  to  $-\infty$  if  $a_3 < 0$ . A cubic polynomial can be continuously either increasing or decreasing, or it can have two bumps. Observe also that such a polynomial always meets the  $x$ -axis, providing a real root either once or three times (see also 1.A.6 in Chapter 1). Later in Section C we will discuss conditions distinguishing these cases.  $\square$

*Functions as trigonometric, exponential or logarithmic, have an elementary role in our zoo of functions, hence we will meet them very often. It will be also useful to know how to treat and evaluate them in Sage, a goal that we quickly figure out for the trigonometric one via our next task.*

**5.A.3. Trigonometric functions via Sage.** Use Sage to obtain the exact values of all the elementary trigonometric functions at the points  $0, \pi/6, \pi/4, \pi/3, \pi/2, 2\pi/3, \pi, 3\pi/2$ , and  $2\pi$ , in case they are defined there.

**Solution.** All the trigonometric functions are built into Sage. We type `sin`, `cos` and `tan` for  $\sin$ ,  $\cos$  and  $\tan = \frac{\sin}{\cos}$ , respectively. Similarly we type `cot`, `sec` and `csc` for the cotangent, secant, and cosecant function. To save your time, we recall that these are the functions defined by  $\cot(x) = \frac{\cos(x)}{\sin(x)} = \frac{1}{\tan(x)}$ ,  $\sec(x) = \frac{1}{\cos(x)}$  and  $\csc(x) = \frac{1}{\sin(x)}$ , respectively.

To obtain the required values in an exact way, you may type `sin(pi/6)`, `cos(3*pi/2)`, and so on.<sup>2</sup> When a function is not defined at a certain point, Sage returns `Infinity`. This is the case for the commands `cot(0)` and `sec(pi/2)`, for example. In this way one may recover the following table:

	sin	cos	tan	cot	sec	csc
0	0	1	0	—	1	—
$\pi/6$	$1/2$	$\sqrt{3}/2$	$\sqrt{3}/3$	$\sqrt{3}$	$2\sqrt{3}/3$	2
$\pi/4$	$\sqrt{2}/2$	$\sqrt{2}/2$	1	1	$\sqrt{2}$	$\sqrt{2}$
$\pi/3$	$\sqrt{3}/2$	$1/2$	$\sqrt{3}$	$\sqrt{3}/3$	2	$2\sqrt{3}/3$
$\pi/2$	1	0	—	0	—	1
$2\pi/3$	$\sqrt{3}/2$	$-1/2$	$-\sqrt{3}$	$-\sqrt{3}/3$	-2	$2\sqrt{3}/3$
$\pi$	0	-1	0	—	-1	—
$3\pi/2$	-1	0	—	0	—	-1
$2\pi$	0	1	0	—	1	—

Notice the inverse trigonometric functions are also built into Sage. Hence we type `arcsin`, `arccos`, `arctan`, `arccot`, `arcsec` and `arccsc` for the inverse of  $\sin$ ,  $\cos$ ,  $\tan$ ,  $\cot$ ,  $\sec$  and  $\csc$ , respectively. You may use Sage to evaluate these functions, or to obtain their graph. For some of them, we will do this later in this chapter.  $\square$

*Let us now focus on polynomial interpolation. Recall by 5.1.3 that given some distinct nodes  $x_0, x_1, \dots, x_n$  there is precisely a polynomial  $P(x)$  of degree not greater than  $n$ ,*

<sup>2</sup>In case you need a decimal approximation, type `N(sin(pi/6))`, etc. An alternative can be the command `n(sin(pi/6))`, though we should be careful not having introduced in the same cell “ $n$ ” as a symbolic variable.

Consider uniqueness. Suppose there are polynomials  $q, q', r$ , and  $r'$ , such that

$$f = q \cdot g + r = q' \cdot g + r'.$$

Then subtraction gives  $0 = (q - q') \cdot g + (r - r')$ .

If  $q = q'$ , then also  $r = r'$ . If  $q \neq q'$ , then the term of highest degree in  $(q - q') \cdot g$  cannot be replicated in  $r - r'$ . This leads to a contradiction. This proves uniqueness.

It remains to prove that  $f$  can always be expressed in the desired form. If  $m > n$ , then  $f = 0 \cdot g + f$  satisfies the requirements. So suppose that  $n \geq m$ . The result is proved by induction on the degree of  $f$ .

If  $f$  is of degree zero, then the statement is trivial. Suppose the statement holds for all polynomials  $f$  of degree less than  $n > 0$ . Put

$$h(x) = f(x) - \frac{a_n}{b_m} x^{n-m} g(x).$$

If  $h(x)$  is the zero polynomial, then  $f$  is of the desired form. Otherwise  $h(x)$  is a polynomial of degree less than that of  $f$  and so  $h$  be written in the desired form as  $h(x) = q \cdot g + r$ . But then

$$f(x) = h(x) + \frac{a_n}{b_m} x^{n-m} g(x) = (q + \frac{a_n}{b_m} x^{n-m})g(x) + r$$

and the proof is complete.  $\square$

If  $f(b)$  equals zero for some element  $b \in \mathbb{K}$ , then the equality  $f(x) = q(x)(x-b) + r$  yields  $0 = f(b) = q(b) \cdot 0 + r$ , and so the constant remainder is  $r = 0$ . Consequently  $f(x) = (x-b)q(x)$ . The value  $b$  is called a *root of the polynomial*  $f$ . The degree of  $q$  is then  $n - 1$ . If  $q$  also has a root, we can continue and in no more than  $n$  steps we arrive at a constant polynomial. It follows that the number of roots of any non-zero polynomial over the field  $\mathbb{K}$  is at most the degree of the polynomial. Hence the following observation:

**Corollary.** *If the field of scalars  $\mathbb{K}$  is infinite, then the polynomials  $f$  and  $g$  are equal as mappings if and only if they are equal as sequences of coefficients.*

**PROOF.** Suppose that  $f = g$ , i.e.  $f - g = 0$ , as a mapping. Then the polynomial  $(f - g)(x)$  has infinitely many roots, which is possible only if it is the zero polynomial.  $\square$

Notice that of course, this statement does not hold for finite fields. A simple counter-example is the polynomial  $x^2 + x$  over  $\mathbb{Z}_2$  which represents a constant zero mapping.

**5.1.3. Interpolation polynomial.** It is often desirable to use an easily computable expression for a function which is given by its values at some given points  $x_0, \dots, x_n$ . Mostly this would be an approximation of an unknown function represented by the finite values only. We look for such polynomials.

If the values were all zeros, we can immediately find a polynomial of degree  $n + 1$ , namely

$$f(x) = (x - x_0)(x - x_1) \dots (x - x_n).$$



which takes a prescribed value  $y_i$  at  $x_i$ , for all  $i = 0, 1, \dots, n$ . This is given by  $P(x) = y_0 \cdot \ell_0(x) + \dots + y_n \cdot \ell_n(x)$ , with

$$\ell_i(x) := \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

for any  $i = 0, 1, \dots, n$ . Observe that  $\ell_i(x_i) = 1$  for all  $i = 0, 1, \dots, n$ , while  $\ell_i(x_j) = 0$  for all  $i \neq j$ .

The polynomial  $P(x)$  is known as the Lagrange interpolation polynomial, while  $\ell_i$  are referred to as the elementary Lagrange polynomials, see 5.1.4 for details. Notice if we have  $y_i = f(x_i)$  for some function  $f$ , then  $P(x)$  is referred to as the Lagrange interpolation polynomial for  $f$ . Let us now illustrate the situation by examples.



**5.A.4. Lagrange interpolation.** Consider the nodes  $x_0 = 0$ ,  $x_1 = 1$ ,  $x_2 = 4$  and the values  $y_0 = 1$ ,  $y_1 = 2$ ,  $y_2 = 4$ . Write down the corresponding Lagrange interpolation polynomial  $P(x)$ , and sketch its graph, together with the graphs of  $\ell_0, \ell_1, \ell_2$ .

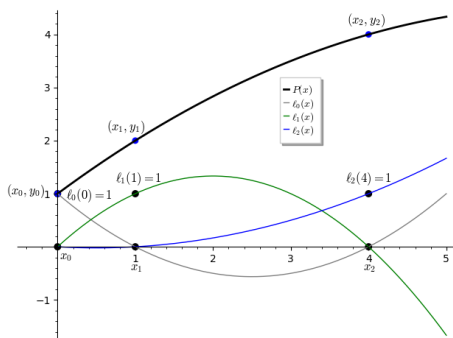
**Solution.** According to 5.1.4 the Lagrange interpolation polynomial will be at most of degree two. We compute  $\ell_0(x) = \frac{(x-1)(x-4)}{4}$ ,  $\ell_1(x) = -\frac{x(x-4)}{3}$  and  $\ell_2(x) = \frac{x(x-1)}{12}$ , and hence  $P(x) = -\frac{1}{12}x^2 + \frac{13}{12}x + 1$ . For an implementation of the Lagrange interpolation, Sage provides an inbuilt method based on the command `lagrange.polynomial`. For example, in our case we can give the block

```
nodes=[(0, 1), (1, 2), (4, 4)]
R=PolynomialRing(QQ, "x")
P=R.lagrange_polynomial(nodes); show(P)
```

Check yourself that this cell prints out the explicit expression of the polynomial  $P(x)$ . Notice also that you can use the `bool` command to confirm Sage's output. This can be done by adding in the cell posed above the following syntax:

```
l0(x)=(x-1)*(x-4)/4; l1(x)=-x*(x-4)/3
l2(x)=x*(x-1)/12
eq=P(x)-l0(x)-2*l1(x)-4*l2(x)==0; eq
bool(eq)
```

As for the required graphs, we present them below (check yourself that the reproduction of this figure in Sage is really enjoyable).

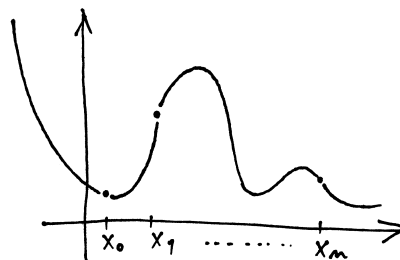


This is zero at these points and only at them. However, there are other polynomials which are zero at the given points. For instance the zero polynomial, which turns out to be the only such polynomial in the vector space of polynomials of degree at most  $n$ . The general situation is analogous:

INTERPOLATION POLYNOMIALS

Let  $\mathbb{K}$  be an infinite field of scalars. An interpolation polynomial  $f$  for the set of (pairwise distinct) points  $x_0, \dots, x_n \in \mathbb{K}$  and given values  $y_0, \dots, y_n \in \mathbb{K}$  is either the zero polynomial, or a polynomial of degree at most  $n$  such that  $f(x_i) = y_i$  for all  $i = 0, 1, \dots, n$ .

**Theorem.** For every set of  $n + 1$  (pairwise distinct) points  $x_0, \dots, x_n \in \mathbb{K}$  and given values  $y_0, \dots, y_n \in \mathbb{K}$ , there is exactly one interpolation polynomial  $f$ .



**PROOF.** If  $f$  and  $g$  are interpolation polynomials with the same defining values, then their difference is a polynomial of degree  $n$  which has at least  $n + 1$  roots, and thus  $f - g = 0$ . This proves uniqueness.

It remains to prove the existence. Label the coefficients of the polynomial  $f$  of degree  $n$ :

$$f = a_n x^n + \dots + a_1 x + a_0.$$

Substituting the desired values leads to a system of  $n + 1$  equations for the same number of unknown coefficients  $a_i$

$$\begin{aligned} a_0 + x_0 a_1 + \dots + (x_0)^n a_n &= y_0 \\ &\vdots \\ a_0 + x_n a_1 + \dots + (x_n)^n a_n &= y_n. \end{aligned}$$

The existence of a solution of this system is easily shown by constructing the polynomial by using the Lagrange polynomials for the given points  $x_0, \dots, x_n$ , introduced in the next paragraph.

However, the proof can be concluded by using only basic knowledge from linear algebra. This system of linear equations has a unique solution if the determinant of its matrix is a non-zero scalar (see 2.3.5 and 2.2.11). The determinant is the *Vandermonde determinant*, which was discussed in the exercise 2.E.21 on page 151.

Since it is verified that for zero right-hand sides, there is exactly one solution, we know that this determinant must be non-zero. □



**5.A.5.** Find a polynomial  $P$  satisfying the following conditions:  $P(2) = 1, P(3) = 0, P(4) = -1, P(5) = 6$ .

**Solution.** Initially one may want to use Sage to plot the given points in  $\mathbb{R}^2$ . We can do this by the cell

```
list_plot({2: 1, 3: 0, 4: -1, 5: 6}, size=30,
         figsize=4, color="black")
```

Now, we can solve the task by two different ways. Four points are given, so we know that there is exactly one polynomial of degree at most three, satisfying the given conditions. Hence we can assume that  $P(x) = a_3x^3 + a_2x^2 + a_1x_1 + a_0$ , for some  $a_0, \dots, a_3 \in \mathbb{R}$ , and in such terms it is easy to see that the system  $\{P(2) = 1, P(3) = 0, P(4) = -1, P(5) = 6\}$  consists of the equations  $a_0 + 2a_1 + 4a_2 + 8a_3 = 1, a_0 + 3a_1 + 9a_2 + 27a_3 = 0, a_0 + 4a_1 + 16a_2 + 64a_3 = -1$ , and  $a_0 + 5a_1 + 25a_2 + 125a_3 = 6$ . Applying the Sage cell

```
var("a0, a1, a2, a3")
eq1=a0+2*a1+4*a2+8*a3-1; eq2=a0+3*a1+9*a2+27*a3
eq3=a0+4*a1+16*a2+64*a3-1
eq4=a0+5*a1+25*a2+125*a3-6
solve([eq1==0, eq2==0, eq3==0, eq4==0], a0, a1, a2, a3)
```

we obtain a unique solution given by

```
[a0 == -29, a1 == (101/3), a2 == -12, a3 == (4/3)]
```

An alternative, more “solid” way to treat this system via Sage is here:

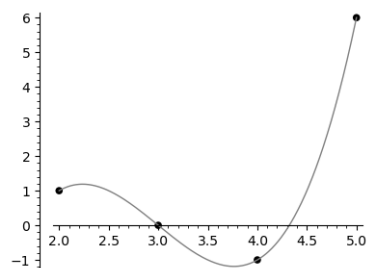
```
a=SR.var("a", 4)
P(x)=a[0]+a[1]*x+a[2]*x^2+a[3]*x^3
solve([P(2)-1==0, P(3)==0, P(4)+1==0, P(5)-6==0], a[0], a[1], a[2], a[3])
```

To summarize, the polynomial  $P$  has the form  $\frac{4}{3}x^3 - 12x^2 + \frac{101}{3}x - 29$ .

Let us now verify this answer by the method described before, based on the Lagrange polynomial:

$$\begin{aligned}
 P(x) &= 1 \cdot \frac{(x-3)(x-4)(x-5)}{(2-3)(2-4)(2-5)} + 0 \cdot (\dots) \\
 &\quad + (-1) \cdot \frac{(x-2)(x-3)(x-5)}{(4-2)(4-3)(4-5)} \\
 &\quad + 6 \cdot \frac{(x-2)(x-3)(x-4)}{(5-2)(5-3)(5-4)} \\
 &= \frac{4}{3}x^3 - 12x^2 + \frac{101}{3}x - 29.
 \end{aligned}$$

Of course, such a computation can be done faster in Sage via the `lagrange.polynomial` command, as in 5.A.4, and we leave to the reader this verification. Below we present the graph of  $P(x)$  together with the given nodes.



Since polynomials are equal as mappings if and only if they are equal as sequences of coefficients, the theorem is proved.  $\square$

**5.1.4. Applications of interpolation.** At first sight, it may seem that real or rational polynomials, that is, polynomial functions  $\mathbb{R} \rightarrow \mathbb{R}$  or  $\mathbb{Q} \rightarrow \mathbb{Q}$ , form a very useful class of functions of one variable. We can arrange for them to attain any set of given values. Moreover, they are easily expressible, so their value at any point can be calculated without difficulties.



However, there are a number of problems when trying to use them in practice.

The first of the problems is to find quickly the polynomial which will interpolate the given data. Solving the aforementioned system of linear equations generally requires time proportional to the cube of the number of given points  $x_i$ . This is unacceptable for large data. We will demonstrate how to overcome this on one popular type of polynomials related to fixed points  $x_0, \dots, x_n$ :

LAGRANGE<sup>1</sup> INTERPOLATION POLYNOMIALS

The *Lagrange interpolation polynomial* is expressed in terms of the *elementary Lagrange polynomials*  $\ell_i$  of degree  $n$  with the properties

$$\ell_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

These polynomials must (up to a constant factor) equal the expressions  $(x-x_0) \dots (x-x_{i-1})(x-x_{i+1}) \dots (x-x_n)$ . So

$$\ell_i(x) = \frac{\prod_{j \neq i} (x-x_j)}{\prod_{j \neq i} (x_i-x_j)}.$$

The desired Lagrange interpolation polynomial is then given by

$$f(x) = y_0\ell_0(x) + y_1\ell_1(x) + \dots + y_n\ell_n(x).$$

Notice that the elementary Lagrange polynomials can be quite easily expressed by means of the derivatives, see the exercise ??.

The usage of Lagrange polynomials is especially efficient when working with different values  $y_i$  for the same set of values  $x_i$ . For in this case, the elementary polynomials  $\ell_i$  are already prepared.

One of the disadvantages of this expression is a large sensitivity to inaccuracies in a computation when the differences of the given values  $x_i$  are small. This is because division by these differences is required.

Another disadvantage (common to all ways of expressing the unique interpolation polynomial) is poor stability of the values of real or rational polynomials outside of the interval containing all its roots.

<sup>1</sup>Joseph-Louis Lagrange (1736-1813) was a famous Italian mathematician and astronomer, who contributed in particular to celestial mechanics. His famous *Mécanique analytique* appeared in 1788. His name appears often even in this elementary textbook.

$\square$

**5.A.6.** Find the Lagrange interpolation polynomial for the function  $f(x) = \frac{1}{4+x^2}$  by dividing the closed interval  $[-1, 1] \subset \mathbb{R}$  into  $n$  equal parts, for  $n = 20$ . Use Sage to present a figure including the graph of  $f$  and that of the interpolation polynomial, along with the interpolated points.

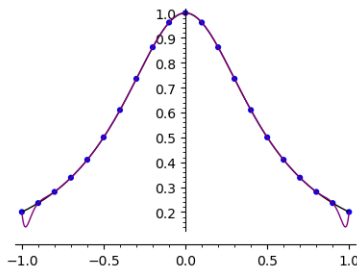
**Solution.** The procedure of dividing  $[-1, 1]$  to  $n$  equal parts gives rise to  $n + 1$  points of the form

$$-1 + \frac{2k}{n}, \quad k = 0, 1, \dots, n.$$

For instance for  $n = 1$  we get the points  $-1, 1$ , for  $n = 2$  we get the points  $-1, 0, 1$ , for  $n = 3$  we get the points  $-1, -1/3, 1/3, 1$ , etc. We can now use Sage and the command `.lagrange_polynomial()`, as follows:

```
f(x)=1/(4+x**2)
fig=plot(f(x), (x, -1, 1), figsize=4, color="black")
show(fig)
R=PolynomialRing(QQ, "x"); n=20
points=[(-1+k*(2/n), f(-1+k*2/n)) for k
        in [0, 1, .., n]]
P(x)=R.lagrange_polynomial(points)
fig+=plot(P(x), (x, -1, 1), color="purple")
fig+=list_plot(points, size=20, figsize=4,
               color="blue"); fig
```

This block includes commands appropriate to produce the plots of  $f$  and of  $P$ , along with the interpolated points. This allows us to illustrate the interpolation, as in the figure below. Notice to obtain the explicit form of  $P(x)$  one should type `show(P(x))`, which we do not present to save some space (type alone in your editor the given code to read  $P(x)$ ).



**5.A.7.** Repeat the task in 5.A.6 by replacing  $f$  with the function  $g(x) = (x/(1 + 8x^2))$ . Next present the graphs of  $g(x)$  and of the Lagrange interpolation polynomial along with the given points, via a programming package of your desire. ○

**5.A.8.** Find a polynomial  $P$  of degree two or less, taking the values  $y_0 = 1, y_1 = -3, y_2 = 4$  at the points  $x_0 = -1, x_1 = 1, x_2 = 2$ , respectively. ○

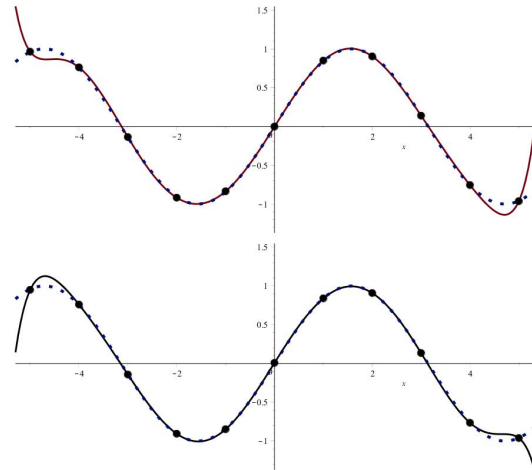
**5.A.9.** Find a polynomial  $P$  of third degree satisfying  $P(0) = 1, P(1) = 0, P(2) = 1$ , and  $P(3) = 10$ . ○

**5.A.10.** Find a polynomial  $P$  satisfying:

- (i)  $P(1 + i) = i, P(2) = 1, P(3) = -i$ ,
- (ii)  $P(1) = i, P(-1) = -i, P(i) = -1$ .

Soon we will develop tools for an exact description of the functions' behaviour. But even without such tools, it is clear that, according to the sign of the coefficient of the term with highest degree, the value of the polynomial will rapidly approach plus or minus infinity as  $x$  increases (or decreases).

However, the above mentioned sign is even not stable under small changes of the defining values  $y_i$ . This is illustrated by the following two diagrams, displaying eleven values of the function  $\sin(x)$  with two different small changes of values. The interpolated function  $\sin(x)$  is the dotted line, the circles are the gently moved values  $y_i$  and the uniquely determined interpolation polynomial is the solid line. While the approximation is quite good inside the interval covering the eleven points, it is very poor at the margins.



There is a rich theory about the interpolation polynomials. If interested, consult the special literature.

**5.1.5. Remark.** Numerical instability caused by the closeness of (some of) the points  $x_i$  is clearly seen in the system of equations from the proof of the Theorem 5.1.3. When solving a system of linear equations, instability is closely related to the size of the determinant of the corresponding matrix. This is the Vandermonde determinant  $V$  in our case.



□ **Lemma.** For any sequence of pairwise distinct scalars  $x_0, \dots, x_n \in \mathbb{K}$ ,

$$V(x_0, \dots, x_n) = \prod_{i>k=0}^n (x_i - x_k).$$

**PROOF.** The proof is by induction on the number of the points  $x_i$ . The result is true for  $n = 1$ . (The problem is completely uninteresting for  $n = 0$ ). Suppose that the result is true for  $n - 1$ , i.e.

$$V(x_0, \dots, x_{n-1}) = \prod_{i>k=0}^{n-1} (x_i - x_k).$$

Consider the values  $x_0, \dots, x_{n-1}$  to be fixed, and vary the value of  $x_n$ . Expand the determinant by the last row (see

**Solution.** A straightforward computation reveals the expression  $P(z) = (-\frac{3}{5} - \frac{4}{5}i)z^2 + (2 + 3i)z - \frac{3}{5} - \frac{14}{5}i$ . We can compute this directly via Sage, as in 5.A.5, although we have to change the field into  $\mathbb{C}$ :

```
R = PolynomialRing(CC, "z")
R.lagrange_polynomial([(1+I, I), (2, 1), (3, -I)])
Sage returns the answer as follows:
(-0.6000000000000000 - 0.8000000000000000*I)*z^2
+ (2.0000000000000000 + 3.0000000000000000*I)*z
- 0.6000000000000000 - 2.8000000000000000*I
```

In the second case, the solution is easier. This is because the conditions are satisfied of the rotation by angle  $\pi/2$  in the complex plane. This means that the polynomial must be of the form  $f(z) = iz$ .  $\square$

*Assuming a bit of experience on elementary functions, we now proceed with tasks highlighting the interplay between elementary functions and polynomial interpolation. Later, in Chapter 7, we will study the so called Chebyshev polynomials, which may enrich this relationship.*

**5.A.11.** Based on Lagrange interpolation, present an approximate polynomial formula of the sine function, using the known values of  $\sin(x)$  at the points  $0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}$ . Next present the graphs of the interpolation polynomial and of  $\sin(x)$  for  $x \in [0, \pi]$ , including in your figure the given nodes.

**Solution.** According to the statement we have the table

$x$	0	$\pi/6$	$\pi/4$	$\pi/3$	$\pi/2$
$\sin(x)$	0	1/2	$\sqrt{2}/2$	$\sqrt{3}/2$	1

To solve our task we can use this table and apply the same technique in Sage, as above. This can be done by the block

```
nodes=[(0, 0), (pi/6, 1/2), (pi/4, sqrt(2)/2),
(pi/3, sqrt(3)/2), (pi/2, 1)]
R=PolynomialRing(RR, "x")
P=R.lagrange_polynomial(nodes); show(P)
```

Sage's answer says that the interpolation polynomial  $P(x)$  is approximately given by  $0.0288x^4 - 0.2043x^3 + 0.0214x^2 + 0.9956x$ .

Now, in order to produce the required graphs together with the given points, one may proceed with the code

```
nodes=[(0, 0), (pi/6, 1/2), (pi/4, sqrt(2)/2),
(pi/3, sqrt(3)/2), (pi/2, 1)]
R=PolynomialRing(RR, "x")
P=R.lagrange_polynomial(nodes)
A=plot(P, 0, pi, color="grey")
B=list_plot({0: 0, pi/6: 1/2, pi/4: sqrt(2)/2,
pi/3: sqrt(3)/2, pi/2: 1},
size=30, figsize=4, color="black")
C=plot(sin(x), 0, pi, color="black")
show(A+B+C, figsize=4)
```

We leave to the reader the implementation of this block.  $\square$

2.2.9). This exhibits the desired determinant as the polynomial

$$(1) \quad V(x_0, \dots, x_n) = (x_n)^n V(x_0, \dots, x_{n-1}) + \text{lower degree terms in } x_n.$$

This is a polynomial of degree  $n$  since its coefficient at  $(x_n)^n$  is non-zero, by the induction hypothesis. Evidently, it vanishes at any point  $x_n = x_i$  for  $i < n$  because in that case, the original determinant contains two identical rows. The polynomial is thus divisible by the expression

$$(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}),$$

which itself is of degree  $n$ . It follows that the Vandermonde determinant (as a polynomial in the variable  $x_n$ ) must, up to a multiplicative constant, be given by

$$V(x_0, \dots, x_n) = c \cdot (x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}).$$

Comparing the coefficients at the highest power in (1) with this expression yields  $c = 1$ , which completes the proof.  $\square$

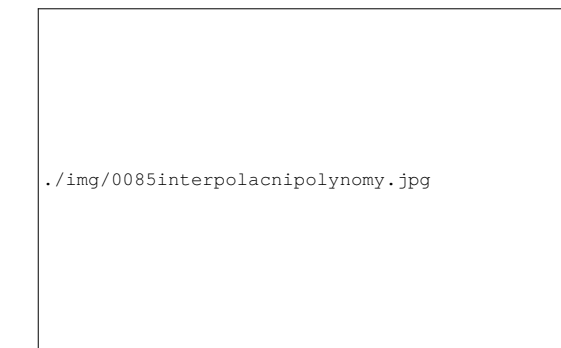
Notice that the value of the determinant is small if the points  $x_i$  are close together.

**5.1.6. Derivatives of polynomials.** The values of polynomials rapidly tend to infinite values as the input variable grows. Hence polynomials are unable to describe periodic events, such as the values of the trigonometric functions. One could say that we will achieve much better results, at least between the points  $x_i$ , if we look not only at the function values, but also at the rate of increase of the function at those points.

For this purpose, we introduce (only intuitively, for the time being) the concept of a *derivative* for polynomials. Again, we can work with real, complex or rational polynomials. The rate of increase of a real-valued polynomial  $f(x)$  at a point  $x \in \mathbb{R}$  should be related to the values

$$(1) \quad \frac{f(x + \delta x) - f(x)}{\delta x},$$

where  $\delta x$  is a small value in  $\mathbb{K}$  expressing the increment of the argument  $x$ . Since we can calculate (over an arbitrary ring)



$$(x + \delta x)^k = x^k + \cdots + \binom{k}{l} x^l (\delta x)^{k-l} + \cdots + (\delta x)^k,$$



The Lagrange interpolation method comes with an interpolation error (and an error bound), which we will briefly analyze in the final section E of this chapter (see 5.E.9, 5.E.10 and 5.E.11).

It is now reasonable to discuss problems where we should also regard the slope of the tangents to our polynomial at the given points. This amounts at employing the derivatives of the polynomials (see 5.1.6 if necessary), and it can be handled by the very same methods as before. Thus next we will meet the so called Hermite interpolation, a method introduced in 5.1.7 (see also 5.1.8).

**5.A.12. Hermite interpolation.** Find a polynomial  $P$  satisfying the following conditions:  $P(1) = 0$ ,  $P'(1) = 1$ ,  $P(2) = 3$ ,  $P'(2) = 3$ .

**Solution.** We provide two methods of finding the solution. 1st approach: The given conditions give rise to four linear equations in the coefficients of  $P$ . If we seek for a polynomial of degree less than four, we get the same number of equations and unknown coefficients. Hence assuming that  $P(x) = a_3x^3 + a_2x^2 + a_1x + a_0$  for some reals  $a_0, \dots, a_3$ , we get the equations

$$\begin{aligned} P(1) &= a_3 + a_2 + a_1 + a_0 = 0, \\ P'(1) &= 3a_3 + 2a_2 + a_1 = 1, \\ P(2) &= 8a_3 + 4a_2 + 2a_1 + a_0 = 3, \\ P'(2) &= 12a_3 + 4a_2 + a_1 = 3. \end{aligned}$$

Solving quickly this linear system by Sage, we get  $P(x) = -2x^3 + 10x^2 - 13x + 5$ .

2nd approach: This is based on Hermite's interpolation, which requires the description of the so-called *fundamental Hermite interpolation polynomials*  $h_i^{(1)}$  and  $h_i^{(2)}$  ( $i = 0, 1$ ), see 5.1.7. By assumption,  $x_0 = 1$ ,  $x_1 = 2$ , and we may set  $y_0 = P(x_0) = 0$ ,  $y_1 = P(x_1) = 3$ ,  $y'_0 = P'(x_0) = 1$  and  $y'_1 = P'(x_1) = 3$ . For an application of the Hermite's interpolation method one needs the function

$$\ell(x) = (x - x_0)(x - x_1) = (x - 1)(x - 2) = x^2 - 3x + 2.$$

Obviously  $\ell'(x) = 2x - 3$ , with  $\ell'(x_0) = -1$  and  $\ell'(x_1) = 1$ . Moreover, the second derivative is a constant function,  $\ell''(x) = 2$ . We also need the elementary Lagrange polynomials induced by the nodes  $x_0, x_1$ :

$$\ell_0(x) = \frac{x - x_1}{x_0 - x_1} = (2 - x), \quad \ell_1(x) = \frac{x - x_0}{x_1 - x_0} = (x - 1).$$

we get for the polynomial  $f(x) = a_nx^n + \dots + a_0$ , the above quotient (1) in the form

$$\begin{aligned} \frac{f(x+\delta x) - f(x)}{\delta x} &= a_n \frac{nx^{n-1}\delta x + \dots + (\delta x)^k}{\delta x} + \dots + a_1 \frac{\delta x}{\delta x} \\ &= na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \dots + a_1 + \delta x(\dots) \end{aligned}$$

where the expression in parentheses in the end of the expression is polynomially dependent on  $\delta x$ . Clearly, for values  $\delta x$  very close to zero, there is a value arbitrarily close to the value in the following definition:

DERIVATIVES OF POLYNOMIALS

The derivative of the polynomial  $f(x) = a_nx^n + \dots + a_0$  with respect to the variable  $x$  is the polynomial

$$f'(x) = na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \dots + a_1.$$

From the definition, it is clear that it is just the value  $f'(x_0)$  of the derivative which gives a good approximation of the polynomial's behaviour near the point  $x_0$ . More precisely, the lines

$$y = \frac{f(x_0 + \delta x) - f(x_0)}{\delta x}(x - x_0) + f(x_0),$$

that is, the secant lines of the graph of the polynomial going through the points  $[x_0, f(x_0)]$  and  $[x_0 + \delta x, f(x_0 + \delta x)]$  approach, as  $\delta x$  decreases, to the line

$$y = f'(x_0)(x - x_0) + f(x_0),$$

which is the "tangent" to the graph of the polynomial  $f$ . This is *linear approximation* to the polynomial  $f$  by its *tangent line*. Exact meaning to all these concepts is given later.

The derivative of polynomials is a linear mapping which, to polynomials of degree at most  $n$ , assigns polynomials of degree at most  $n - 1$ .

Iterating this procedure, there are the second derivative  $f''$ , the third derivative  $f^{(3)}$ , and generally after  $k$ -tuple iteration, the polynomial  $f^{(k)}$  of degree  $n - k$ . Thus the  $(n+1)$ -st derivative is the zero polynomial. This linear mapping is an example of cyclic nilpotent mappings, which are more thoroughly examined in paragraph 3.4.10.

The derivative behaves well also with respect to the multiplication of polynomials. A straightforward combinatorial check reveals the *derivation property* or *Leibniz rule* for this linear operator :

$$(f(x) \cdot g(x))' = f'(x) \cdot g(x) + f(x) \cdot g'(x).$$

Actually this is a purely algebraic result (which holds over any ring of scalars!) and you may either check it yourself or consult the formal proof in 12.3.7.

**5.1.7. Hermite's interpolation problem.** Consider  $m + 1$  pairwise distinct real numbers  $x_0, \dots, x_m$ , i.e.  $x_i \neq x_j$  for all  $i \neq j$ . It is desired to place polynomials through given values at these points, but to determine also the first derivatives of the interpolating polynomial in these points. Set  $y_i$  and  $y'_i$  for all  $i$ .



Now, according to the formulas given in 5.1.7 we have

$$\begin{aligned}
 h_0^{(1)}(x) &:= \left(1 - \frac{\ell''(x_0)}{\ell'(x_0)}(x - x_0)\right) (\ell_0(x))^2 \\
 &= (2x - 1)(x - 2)^2, \\
 h_1^{(1)}(x) &:= \left(1 - \frac{\ell''(x_1)}{\ell'(x_1)}(x - x_1)\right) (\ell_1(x))^2 \\
 &= (5 - 2x)(x - 1)^2, \\
 h_0^{(2)}(x) &:= (x - x_0)(\ell_0(x))^2 = (x - 1)(x - 2)^2, \\
 h_1^{(2)}(x) &:= (x - x_1)(\ell_1(x))^2 = (x - 2)(x - 1)^2.
 \end{aligned}$$

The polynomial  $P(x)$  is then given by the expression  $y_0 \cdot h_0^{(1)}(x) + y_1 \cdot h_1^{(1)}(x) + y'_0 \cdot h_0^{(2)}(x) + y'_1 \cdot h_1^{(2)}(x)$ , and we arrive to the same expression as above:


$$\begin{aligned}
 P(x) &= 0 \cdot h_0^{(1)}(x) + 3 \cdot h_1^{(1)}(x) + 1 \cdot h_0^{(2)}(x) \\
 &\quad + 3 \cdot h_1^{(2)}(x) = -2x^3 + 10x^2 - 13x + 5.
 \end{aligned}$$

In fact, one could avoid to compute explicitly  $h_0^{(1)}$ , since  $y_0 = 0$  by assumption. Notice also that our final computation can be verified quickly in Sage, via the block

```

h01(x)=(2*x-1)*(x-2)^2; h11(x)=(5-2*x)*(x-1)^2
h02(x)=(x-1)*(x-2)^2; h12(x)=(x-2)*(x-1)^2
y0=0; y1=3; dy0=1; dy1=3
P(x)=y0*h01(x)+y1*h11(x)+dy0*h02(x)+dy1*h12(x)
show(P(x).expand())
    
```

□

 *The next exercises on Hermite interpolation are helpful for a better understanding of the method but also for improving your computational skills (we suggest to treat these tasks with the help of Sage, as well). One can find further problems related to Hermite interpolation also in Section E, see 5.E.1.*

**5.A.13.** Determine the Hermite interpolation polynomial  $Q$  satisfying  $Q(-1) = -9$ ,  $Q(1) = -1$ ,  $Q'(-1) = 10$ ,  $Q'(1) = 2$ . ○

**5.A.14.** Replace the function  $f$  with a Hermite polynomial, having as initial data the following values of  $f$ :

$x_i$	-1	1	2
$f(x_i)$	4	-4	-8
$f'(x_i)$	8	-8	11

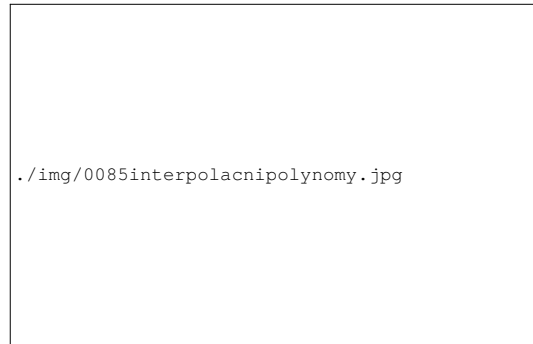
○

*Let us now focus on splines, a notion introduced in 5.1.9. Shortly, we look for (cubic) polynomials on intervals, prescribing the values and requesting the first and second derivatives to be shared in the boundary points. Splines are very popular in environmental research for image processing, where smooth interpolation becomes essential, e.g., for computational animation and image scaling.*

*However, a formal description of the spline through a given data can be often a tedious task. For instance, if we are*

A polynomial  $f$  is wanted which will satisfy these conditions on the values and derivatives.

As in the case of interpolating the values only, we obtain the following system of  $2(m+1)$  equations for the coefficients of the polynomial  $f(x) = a_n x^n + \dots + a_0$ :



$$\begin{aligned}
 a_0 + x_0 a_1 + \dots + (x_0)^n a_n &= y_0 \\
 &\vdots \\
 a_0 + x_m a_1 + \dots + (x_m)^n a_n &= y_m \\
 a_1 + 2x_0 a_2 + \dots + n(x_0)^{n-1} a_n &= y'_0 \\
 &\vdots \\
 a_1 + 2x_m a_2 + \dots + n(x_m)^{n-1} a_n &= y'_m.
 \end{aligned}$$

We could verify that with the choice  $n = 2m + 1$ , the determinant of this system is non-zero, and thus there is exactly one solution.

The polynomial  $f$  can be constructed immediately. Simply create a set of polynomials with values 0 or 1 respectively for the derivatives and the values, in order to express the desired values as the linear combination. We sketch briefly, how to construct them now, leaving the details to the reader.

The elementary Lagrange polynomials serve well for this purpose. The derivative of  $f(x) = (\ell_i(x))^2$  is  $2\ell'_i(x)\ell_i(x)$  and thus all  $x_j$  are roots of this polynomial, except for  $j = i$ . Similarly for the derivative  $f'(x)$ . But a polynomial of degree  $2m + 1$  is wanted. So we consider rather  $g(x) = (x - x_i)f(x)$ . Now the values will be all zero, while the derivative  $g'(x) = f(x) + (x - x_i)f'(x)$  has the required properties too. Thus we take  $h_i^{(2)}(x) = (x - x_i)(\ell_i(x))^2$ . This is called the *fundamental Hermitian polynomial*<sup>2</sup> of the second type.

Finally we look for a polynomial which has zero derivatives at all points  $x_i$  with the same values as  $\ell_i$  at the given points  $x_i$ . We can apply a very similar trick. Look for polynomials of the form  $h_i^{(1)}(x) = (1 - a(x - x_i))(\ell_i(x))^2$ . All  $x_j$  will be roots of this polynomial, except for  $j = i$  where

<sup>2</sup>Charles Hermite (1822-1901) was a Frenchman active in many areas of Mathematics. His name is mostly linked to the Hermitian operators and matrices, cf. 3.4.6.

given  $n \geq 2$  points and values in them, then we need to solve  $4n - 4$  linear equations. Fortunately, using computers we may establish algorithms which lead to much faster computer implementations. In order to illustrate this situation, below we proceed by combining the appropriate syntax of Sage. But let us first present a more basic example.

**5.A.15. Splines.** Find a natural cubic spline  $S$  satisfying  $S(-1) = 0$ ,  $S(0) = 1$ ,  $S(1) = 0$ , and present its graph for  $x$  in the closed interval  $[-1, 1]$ .

**Solution.** We have three nodes and the intervals  $[x_0, x_1] = [-1, 0]$  and  $[x_1, x_2] = [0, 1]$ , that is,  $n = 2$  in terms of the paragraph 5.1.9. Thus the spline that we are looking for consists of two cubic polynomials, say  $S_1(x) = ax^3 + bx^2 + cx + d$ , and  $S_2(x) = ex^3 + fx^2 + gx + h$ , with domain the intervals  $[-1, 0]$  and  $[0, 1]$ , respectively, with  $a, b, \dots, h \in \mathbb{R}$ . Applying the definition of cubic splines for  $n = 2$ , one obtains six linear equations in terms of  $a, b, \dots, h$ :

$$S_1(-1) = 0, S_2(0) = 1, S_1(0) = 1, S_2(1) = 0, \\ S_1'(0) = S_2'(0), S_1''(0) = S_2''(0).$$

In addition,  $S$  must be a natural spline, and this corresponds to the vanishing of the second derivatives of  $S_1 = ax^3 + bx^2 + cx + d$  and  $S_2 = ex^3 + fx^2 + gx + h$  at the points  $-1$  and  $1$ , respectively:  $S_1''(-1) = 0$  and  $S_2''(1) = 0$ . Hence we arrive to a system of eight equations, which in fact we can reduce in the following way. Thanks to the given value at  $0$ , we know that the absolute coefficients of both the polynomials are  $1$ . The resulting spline has to be symmetric along the  $y$  axis, otherwise we would get two splines satisfying the condition by the reflection along the axis. Here we are based on the fact that a natural cubic spline is unique. Thus, the only possibility for the common values for the first derivatives of  $S_1$  and  $S_2$  at zero is zero, further the second derivatives in zero have to agree, that is  $b = d$ , and the symmetry gives also  $a = -c$ .

So far we have  $S_1(x) = ax^3 + bx^2 + 1$  and  $S_2(x) = -ax^3 + bx^2 + 1$ . Now the conditions  $S_1(-1) = 0$  and  $S_1'(-1) = 0$  correspond to the equations  $-a + b + 1 = 0$  and  $-6a + 2b = 0$ , respectively. Solving this system we obtain  $S_1(x) = -\frac{1}{2}x^3 - \frac{3}{2}x^2 + 1$ ,  $S_2(x) = \frac{1}{2}x^3 - \frac{3}{2}x^2 + 1$ . Altogether,

$$S(x) = \begin{cases} -\frac{1}{2}x^3 - \frac{3}{2}x^2 + 1, & \text{if } x \in [-1, 0], \\ \frac{1}{2}x^3 - \frac{3}{2}x^2 + 1, & \text{if } x \in [0, 1]. \end{cases}$$

For the graph of the spline via Sage, we obtain the figure presented below (we used blue colour for  $S_1$  and green for  $S_2$ )

the value is  $1$ . The derivative is

$$(h_i^{(1)}(x))' = -a(\ell_i(x))^2 + (1 - a(x - x_i))2\ell_i(x)\ell_i'(x).$$

All  $x_j, j \neq i$ , are roots of  $\ell_i(x)$ , thus they are also roots of this polynomial. Finally, at the point  $x_i$  we want  $0 = -a + 2\ell_i'(x_i)$ . Thus we choose  $a = 2\ell_i'(x_i)$ .

The combinatorial check that  $2\ell_i'(x_i) = \frac{\ell''(x_i)}{\ell'(x_i)}$  is left to the reader. We summarize:

#### HERMITE'S 1ST ORDER INTERPOLATION POLYNOMIAL

The *fundamental Hermite polynomials* are defined as follows:

$$h_i^{(1)}(x) = \left(1 - \frac{\ell''(x_i)}{\ell'(x_i)}(x - x_i)\right) (\ell_i(x))^2 \\ h_i^{(2)}(x) = (x - x_i) (\ell_i(x))^2,$$

where  $\ell(x) = \prod_{i=0}^m (x - x_i)$ ,  $\ell_i(x)$  is the elementary Lagrange polynomial. These polynomials satisfy:

$$h_i^{(1)}(x_j) = \delta_i^j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \\ (h_i^{(1)})'(x_j) = 0 \\ h_i^{(2)}(x_j) = 0 \\ (h_i^{(2)})'(x_j) = \delta_i^j.$$

The *Hermite's interpolation polynomial* is given by the expression

$$f(x) = \sum_{i=0}^m (y_i h_i^{(1)}(x) + y_i' h_i^{(2)}(x)).$$

**5.1.8. Examples of Hermite's polynomials.** The simplest example is the one of prescribing the value and the derivative at one point. This determines a polynomial of degree one

$$f(x) = f(x_0) + f'(x_0)(x - x_0).$$

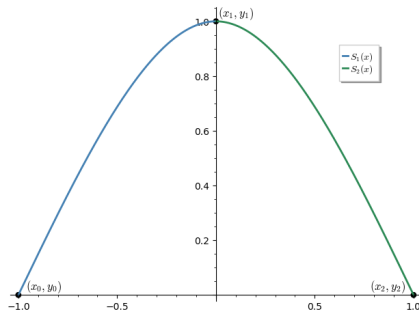
This is exactly the equation of the straight line given by the value and slope at the point  $x_0$ . When we set the values and the derivatives at two points, i.e.  $y_0 = f(x_0)$ ,  $y_0' = f'(x_0)$ ,  $y_1 = f(x_1)$ ,  $y_1' = f'(x_1)$  for two distinct points  $x_i$ , we still obtain an easily computable problem.

Consider the simple case when  $x_0 = 0$ ,  $x_1 = 1$ . Then the matrix of the system and its inverse is

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 3 & 2 & 1 & 0 \end{pmatrix}, A^{-1} = \begin{pmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

The multiplication  $A^{-1} \cdot (y_0, y_1, y_0', y_1')^T$  gives the vector  $(a_3, a_2, a_1, a_0)^T$  of coefficients of the polynomial  $f$ , i.e.

$$f(x) = (2y_0 - 2y_1 + y_0' + y_1')x^3 \\ + (-3y_0 + 3y_1 - 2y_0' - y_1')x^2 + y_0'x + y_0.$$



**5.A.16. Splines via Sage.** Splines can be treated in Sage via the command `spline`. For instance, a graphical implementation of the previous example in Sage relies on the given interpolated points and goes as follows:

```
pts=[(-1, 0), (0, 1), (1, 0)]; S=spline(pts)
a=plot(S, -1, 1, color="darkgray", figsize=4)
show(points(pts)+a)
```

Verify yourself that for the spline in question this syntax gives us the same graph (so this provides a graphical verification of our previous computation).

Changing the points of the spline causes of course the spline to be recomputed. Notice also that this syntax does not return the explicit form of  $S$ , as a cubic polynomial. However, it allows us to compute the values of the spline, e.g., by adding the syntax

```
show(S(0.4)); show(S(0.5)); show(S(0.6))
```

This returns the values of  $S$  at the chosen points, and one can easily verify that these values coincide with the values that we obtain by applying the given solution in 5.A.15 at the same points. Notice however that we cannot compute the value of the spline at points that are not included in between the given interpolated points.

**5.A.17.** Find a (cubic) spline  $S$  which satisfies  $S(-1) = 0$ ,  $S(0) = 1$ ,  $S(1) = 0$ ,  $S'(-1) = -1$ ,  $S'(1) = 1$ . Hint: Apply the same trick as in 5.A.15. ○

**5.A.18.** Interpolate the function  $f(x) = e^{x^2} - e$  on the interval  $[-6/5, 6/5]$  by the (unique) natural cubic spline  $S$  corresponding to the partition  $x_0 = -1.2$ ,  $x_1 = -1$ ,  $x_2 = 0$ ,  $x_3 = 1$ ,  $x_4 = 1.2$ . Next, using Sage plot the functions  $f$  and  $S$  (together with the interpolated nodes), and based on the “`spline`” method described in 5.A.16, verify your result.

**Solution.** By assumption we have five nodes,  $x_0 = -1.2$ ,  $x_1 = -1$ ,  $x_2 = 0$ ,  $x_3 = 1$ ,  $x_4 = 1.2$ , and so four intervals  $[x_0, x_1]$ ,  $[x_1, x_2]$ ,  $[x_2, x_3]$ ,  $[x_3, x_4]$ . Thus, in terms of the definition of cubic splines given in paragraph 5.1.9 we have

**5.1.9. Spline interpolation.** We can prescribe any finite number of derivatives at the particular points and a convenient choice for the upper bound on the degree of the desired polynomial, leading to a unique interpolation. Unfortunately, these interpolations do not solve the problems mentioned already – complexity of the computations and instability. However, a smarter usage of derivatives allows an improvement.



As we have seen, small local changes of the values may dramatically affect the overall changes of the behaviour of the resulting polynomial. In particular, this may happen outside of the interval covered by the points  $x_i$ . We try gluing together polynomial pieces of low degree.

The simplest possibility is to interpolate each pair of adjacent points by a polynomial of degree at most one. This corresponds either to the interpolation by values with two points, or by guessing the slope and employing Hermite’s first order interpolation at a single point. This is a common way of displaying data. This means that derivatives will be constant on each of the segments, with a discontinuous ‘jump’ at the given points. There is no freedom for improvements.

A more sophisticated method is to prescribe the value and the derivative at each point. We have then four values for each pair of neighbouring points. As seen earlier, this uniquely determines Hermite’s polynomial of degree three. This polynomial can then be used for all the values of the input variable between the distinguished points  $x_0 < x_1$ . Such a piece-wise polynomial approximation has the property that the first derivatives will be compatible (equal) at the meeting points  $x_i$  in the interval  $[x_0, x_1]$ .

In practice, mere compatibility of the first derivatives is often insufficient. Consider for instance railway tracks, where the second derivative corresponds to acceleration. Discontinuous jumps would be very undesirable for the second derivative. So instead of requiring fixed values of the first derivatives, we request equality of both first and second derivatives at adjacent pieces of the cubic polynomials, as well as fixing the values at the given points. This requirement yields the same number of equations and unknowns, and so the problem is solvable similarly to the 1st order Hermite interpolation problem:

CUBIC SPLINES

Let  $x_0 < x_1 < \dots < x_n$  be real values at which the required values  $y_0, \dots, y_n$  are given. A *cubic interpolation spline* for this assignment is a function  $S : \mathbb{R} \rightarrow \mathbb{R}$  which satisfies the following conditions:

- the restrictions of  $S$  on the intervals  $[x_{i-1}, x_i]$  are polynomials  $S_i$  of degree at most three, for all  $i = 1, \dots, n$
- $S_i(x_{i-1}) = y_{i-1}$  and  $S_i(x_i) = y_i$  for all  $i = 1, \dots, n$ ,
- $S'_i(x_i) = S'_{i+1}(x_i)$  for all  $i = 1, \dots, n - 1$ ,
- $S''_i(x_i) = S''_{i+1}(x_i)$  for all  $i = 1, \dots, n - 1$ .

$n = 4$ . This means that the spline  $S(x)$  has the form

$$\begin{cases} S_1(x) = a_3x^3 + a_2x^2 + a_1x + a_0, & x \in [-\frac{6}{5}, -1], \\ S_2(x) = b_3x^3 + b_2x^2 + b_1x + b_0, & x \in [-1, 0], \\ S_3(x) = c_3x^3 + c_2x^2 + c_1x + c_0, & x \in [0, 1], \\ S_4(x) = d_3x^3 + d_2x^2 + d_1x + d_0, & x \in [1, \frac{6}{5}], \end{cases}$$

with  $a_i, b_i, c_i, d_i \in \mathbb{R}$  for all  $i = 0, 1, 2, 3$ . In our terms, we also have  $y_0 = f(x_0) \approx 1.502$ ,  $y_1 = f(x_1) = 0$ ,  $y_2 = f(x_2) \approx -1.718$ ,  $y_3 = f(x_3) = 0$ ,  $y_4 = f(x_4) \approx 1.502$ . Let us summarize the sixteen conditions that determine  $S$ :

$$\begin{aligned} S_1(x_0) = y_0 &\Leftrightarrow -(\frac{6}{5})^3 a_3 + (\frac{6}{5})^2 a_2 - \frac{6}{5} a_1 + a_0 = 1.502, \\ S_2(x_1) = y_1 &\Leftrightarrow -b_3 + b_2 - b_1 + b_0 = 0, \\ S_3(x_2) = y_2 &\Leftrightarrow c_0 = -1.718, \\ S_4(x_3) = y_3 &\Leftrightarrow d_3 + d_2 + d_1 + d_0 = 0, \\ S_1(x_1) = y_1 &\Leftrightarrow -a_3 + a_2 - a_1 + a_0 = 0, \\ S_2(x_2) = y_2 &\Leftrightarrow b_0 = -1.718, \\ S_3(x_3) = y_3 &\Leftrightarrow c_3 + c_2 + c_1 + c_0 = 0, \\ S_4(x_4) = y_4 &\Leftrightarrow (\frac{6}{5})^3 d_3 + (\frac{6}{5})^2 d_2 + \frac{6}{5} d_1 + d_0 = 1.502, \\ S'_1(x_1) = S'_2(x_1) &\Leftrightarrow 3a_3 - 2a_2 + a_1 = 3b_3 - 2b_2 + b_1, \\ S'_2(x_2) = S'_3(x_2) &\Leftrightarrow b_1 = c_1, \\ S'_3(x_3) = S'_4(x_3) &\Leftrightarrow 3c_3 + 2c_2 + c_1 = 3d_3 + 2d_2 + d_1, \\ S''_1(x_1) = S''_2(x_1) &\Leftrightarrow -6a_3 + 2a_2 = -6b_3 + 2b_2, \\ S''_2(x_2) = S''_3(x_2) &\Leftrightarrow b_2 = c_2, \\ S''_3(x_3) = S''_4(x_3) &\Leftrightarrow 6c_3 + 2c_2 = 6d_3 + 2d_2, \\ S''_1(x_0) = 0 &\Leftrightarrow -(\frac{36}{5})a_3 + 2a_2 = 0, \\ S''_4(x_4) = 0 &\Leftrightarrow (\frac{36}{5})d_3 + 2d_2 = 0. \end{aligned}$$

Solving in Sage this system of linear equations we get a unique solution given by

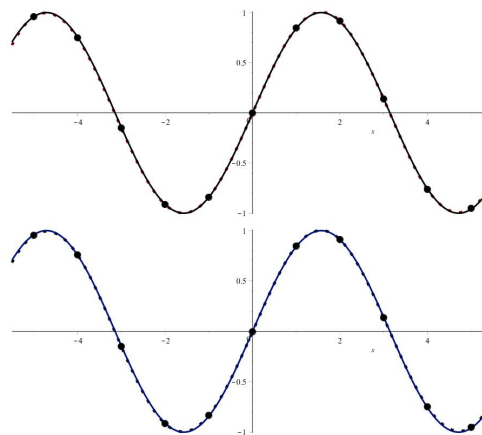
$$\begin{aligned} S_1(x) &= \frac{4933}{380}x^3 + \frac{44397}{950}x^2 + \frac{228243}{4750}x + \frac{135841}{9500}, \\ S_2(x) &= -\frac{28837}{9500}x^3 - \frac{3129}{2375}x^2 - \frac{859}{500}, \\ S_3(x) &= \frac{28837}{9500}x^3 - \frac{3129}{2375}x^2 - \frac{859}{500}, \\ S_4(x) &= -\frac{4933}{380}x^3 + \frac{44397}{950}x^2 - \frac{228243}{4750}x + \frac{135841}{9500}. \end{aligned}$$

Hence the plot of  $S(x)$  in  $[-1.2, 1.2]$  is given by

The cubic spline<sup>3</sup> for  $n + 1$  points consists of  $n$  cubic polynomials. There are  $4n$  free parameters (the first condition from the definition). The other conditions then yield  $2n + (n - 1) + (n - 1)$  more equalities. Two parameters remain free. The values of the derivatives at the marginal points may be prescribed explicitly (the *complete spline*), or the second derivatives can be set to zero (the *natural spline*).

Unfortunately, the computation of the whole spline is not as easy as with the independent computations of Hermite's cubic polynomials because the data mingles between adjacent intervals. However, ordering the variables and equations properly, gives a matrix of the system such that all of its non-zero elements appear only on three diagonals. These matrices are nice enough to be solved in a time proportional to the number of points, using a suitable numerical method. The results are stunning.

For comparison, look at the interpolation of the same data as in the case of the Lagrange polynomial, now using splines. The spline is the solid line, the interpolated function is again the dotted line.



Although the diagrams look nearly identical, the data is different.

## 2. Real numbers and limit processes

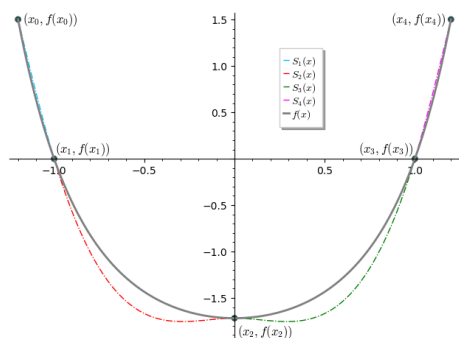
Polynomials and splines do not supply a sufficiently large stock of functions to express many dependencies.

Actually, the first problem to solve is how to define the values of more general functions at all. In principle, all we can get with a finite number of multiplications and additions is polynomial functions. Perhaps division by polynomial quantities, and some efficient manipulation with rational numbers can be done. However, we cannot restrict ourselves to rational numbers. For instance,  $\sqrt{2}$  is not a rational number.

Thus the first step is a thorough introduction to limit processes. We define precisely what it means for a sequence of numbers to approach another number.

<sup>3</sup>The name comes from the name of an elastic ruler used by engineers to draw smooth curve interpolation points. In fact, the requirement on the equality of the first and second derivatives is a good model for natural elasticity behaviour.





Using the command `spline` one can obtain a graphical verification, as before. Although the code is given below, we leave to the reader to check the result alone.

```
f(x)=e^(x^2)-e
plf=plot(f, x, -1.2, 1.2, color="black")
pts=[(-1.2, f(-1.2)), (-1, f(-1)), (0, f(0)),
(1, f(1)), (1.2, f(1.2))]
S=spline(pts)
sps=plot(S, -1.2, 1.2, color="steelblue")
show(points(pts, size=40)+sps+plf)
```

□

**5.A.19.** Without calculation, construct the natural cubic interpolation spline for the points  $x_0 = -1$ ,  $x_1 = 0$  and  $x_2 = 2$  and the value  $y_0 = y_1 = y_2 = 1$  at these points.

**Solution.** The natural splines request the second derivative in the outer boundary points to vanish. Thus, the constant spline  $S_1(x) \equiv 1$ ,  $S_2(x) \equiv 1$  satisfies all conditions and thus it must be the unique solution. □

**5.A.20.** Construct the natural cubic interpolation spline for the function  $f(x) = 1/(1 + x^2)$  selecting the nodes  $x_0 = 0$ ,  $x_1 = 1$ , and  $x_2 = 3$ . ○

*Addition problems concerning polynomial interpolation can be found in the final section E, see also Chapter 6.*

## B. Real numbers and limit processes

*In this section we treat limits of sequences and of functions. We will also discuss some basic topological notions about subsets of the real line  $\mathbb{R}$  or the complex plane  $\mathbb{C}$ . In this way we will establish the foundations of calculus, but also become familiar with elementary ideas from mathematical analysis, that we will revise finally later, in Chapter 7.*

*Functions and limits are some of the most fundamental concepts in mathematics and are ideas that came of age in the 17th century.<sup>3</sup> The French mathematician Augustin-Louis*

<sup>3</sup> The French mathematician Pierre de Fermat (1607-1665) was one of the first realizing the importance of limits. When later Newton and Leibniz invented the calculus they don't use "delta-epsilon" proofs, and it took more than a century to develop them. From this perspective, it should be no wonder that often modern students find an introduction to calculus difficult. Hence you should not worry much whenever some of the notions that we discuss

An important property of polynomials is the “continuous” dependency of their values on the input variable. We expect intuitively that if  $x$  is changed by a little, then the value of  $f(x)$  also changes only a little. This behaviour is not possessed by piece-wise constant functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  near the “jump discontinuities”. For instance, the *Heaviside function*<sup>4</sup>

$$f(x) = \begin{cases} 0 & \text{for all } x < 0, \\ 1/2 & \text{for } x = 0, \\ 1 & \text{for all } x > 0 \end{cases}$$

has this type of “discontinuity” for  $x = 0$ .

We formalize these intuitive statements.

**5.2.1. Real numbers.** We have dealt with algebraic properties of real numbers, summarized by the claim that  $\mathbb{R}$  is a field. However, we have also used the relation of the standard (total) ordering of the real numbers, denoted by “ $\leq$ ”. See the paragraph 1.6.3 on the page 43.

The properties (axioms) of the real numbers, including the connections between the relations and other operations, are enumerated in the following table. The bars indicate how the axioms guarantee that the real numbers form an abelian (commutative) group with respect to addition, that  $\mathbb{R} \setminus \{0\}$  is an abelian group with respect to multiplication, that  $\mathbb{R}$  is a field, that the set  $\mathbb{R}$  together with the operations  $+$ ,  $\cdot$  and the order relation is an *ordered field*. The last axiom can be considered as claiming that  $\mathbb{R}$  is “sufficiently dense”.

<sup>4</sup> Oliver Heaviside was an unconventional English electrical engineer (1850-1925) with an innovative and very original approach to practical mathematical modelling. His famous sayings include “*Mathematics is an experimental science, and definitions do not come first, but later on*”. Or, defending his incomplete argumentation, “*I do not refuse my dinner simply because I do not understand the process of digestion*”. Is this suggestive of the methodology of this textbook?

Cauchy (1789-1867) was probably the first who put the calculus in a rigorous basis. To be more specific, Cauchy introduced the “delta-epsilon” notation commonly encountered in the definition of limits. While such proofs are presented here, we have chosen to de-emphasize them after a certain point. This is because many calculus problems can be approached in various alternative ways. Additionally, this approach allows us to conserve space, which is then utilized to implement solutions using Sage

Our initial tasks involve the concepts of the supremum ( $\sup A$ ) and the infimum ( $\inf A$ ) of a subset  $A$  of real or complex numbers, which are elementary notions discussed in 5.2.1. These represent the smallest upper bound and the largest lower bound of  $A$ , respectively. They always exist when  $A$  is a bounded subset, but they do not necessarily belong to  $A$ , as they can be either limit points or isolated points of  $A$ , see 5.2.5 and 5.2.7.

5.B.1. Find  $\sup A$  and  $\inf A$ , if they exist, for

$$A = \left\{ n + \frac{(-1)^n}{n} : n \in \mathbb{N}^* \right\} \subset \mathbb{R},$$

where as usual we set  $\mathbb{N}^* = \mathbb{N} \setminus \{0\} = \mathbb{Z}_+$ .

5.B.2. Determine the infima and suprema of the sets

$$B = \left\{ \frac{(-1)^n}{n^2} : n \in \mathbb{Z}_+ \right\}, \quad C = (-9, 9) \cap \mathbb{Q},$$

$$X = \left\{ -\frac{1}{n} : n \in \mathbb{Z}_+ \right\}, \quad Y = (0, 2] \cup [3, 5] \setminus \{4\},$$

and look whether they belong to them or not.



Sequence are functions defined on the set of natural numbers. Next we are interesting in the behaviour of such functions as  $n$  tends to infinity, see also the discussion in 5.2.3, 5.2.9 and 5.2.10. To begin with, recall that for any subset  $A$  of reals or complex numbers we can assign the notion of “limit points” of  $A$ . These are points  $x$  for which each of their  $\varepsilon$ -neighbourhoods  $\mathcal{O}_\varepsilon(x) = \{z : |z - x| < \varepsilon\}$  contains at least one point from  $A$  other than  $x$ . Be aware that limit points may not belong to  $A$ . Having in hand this definition of limit points we can define a “convergent sequence”  $(x_n)$  of real or complex numbers, as a sequence approaching its only limit point  $x$ , as  $n$  tends to infinity, see 5.2.3. In this case we say that  $(x_n)$  converges to  $x$  (or that tends to  $x$ ), and write  $x_n \rightarrow x$ , or  $\lim_{n \rightarrow \infty} x_n = x$ . When the limit does not exist as a real, or complex number, or is  $\pm\infty$ , the sequence  $(x_n)$  is called divergent, see also below.

Notice that in our definition of limits points the neighbourhood  $\mathcal{O}_\varepsilon(x)$  is either an open interval in  $\mathbb{R}$ , or an open disc in  $\mathbb{C}$ . The related topological notion of open sets is introduced in 5.2.7, see also 5.2.5. To make the description in this

below seem difficult. To master with limits, it takes a little while. Once it is mastered, however, it becomes easier to understand differentiation and integration.

AXIOMS OF THE REAL NUMBERS

- (R1)  $(a + b) + c = a + (b + c)$ , for all  $a, b, c \in \mathbb{R}$
- (R2)  $a + b = b + a$ , for all  $a, b \in \mathbb{R}$
- (R3) there is an element  $0 \in \mathbb{R}$  such that for all  $a \in \mathbb{R}$ ,  $a + 0 = a$
- (R4) for all  $a \in \mathbb{R}$ , there is an additive inverse  $(-a) \in \mathbb{R}$  such that  $a + (-a) = 0$
- (R5)  $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ , for all  $a, b, c \in \mathbb{R}$
- (R6)  $a \cdot b = b \cdot a$  for all  $a, b \in \mathbb{R}$
- (R7) there is an element  $1 \in \mathbb{R}$ ,  $1 \neq 0$ , such that for all  $a \in \mathbb{R}$ ,  $1 \cdot a = a$
- (R8) for all  $a \in \mathbb{R}$ ,  $a \neq 0$ , there is a multiplicative inverse  $a^{-1} \in \mathbb{R}$  such that  $a \cdot a^{-1} = 1$
- (R9)  $a \cdot (b + c) = a \cdot b + a \cdot c$ , for all  $a, b, c \in \mathbb{R}$
- (R10) the relation  $\leq$  is a total order, i.e. reflexive, anti-symmetric, transitive, and total on  $\mathbb{R}$
- (R11) for all  $a, b, c \in \mathbb{R}$ ,  $a \leq b$  implies  $a + c \leq b + c$
- (R12) for all  $a, b \in \mathbb{R}$ ,  $a > 0$  and  $b > 0$  implies  $a \cdot b > 0$
- (R13) every non-empty set  $A \subset \mathbb{R}$  which has an upper bound has a least upper bound.

The concept of a least upper bound from axiom (R13), also called the *supremum*, is very important. It makes sense for any partially ordered set. This is a set with a (not necessarily total) ordering relation. Recall that an ordering relation is a binary relation on a set which is reflexive, antisymmetric, and transitive; see the paragraph 1.6.3.

SUPREMUM AND INFIMUM

Consider a subset  $A \subset B$  in a partially ordered set  $B$ . An *upper bound* of the set  $A$  is any element  $b \in B$  such that  $b \geq a$  for all  $a \in A$ .

Similarly, a *lower bound* of the set  $A$  is an element  $b \in B$  such that  $b \leq a$  for all  $a \in A$ .

The least upper bound of the set  $A$ , if it exists, is called its *supremum* and it is denoted by  $\sup A$ . Similarly, the greatest lower bound, if it exists, is called the *infimum* and it is denoted by  $\inf A$ .

Thus, the last axiom (R13) from the table of properties of the real numbers can be reformulated as follows: *Every non-empty bounded set  $A$  of real numbers has a supremum. This means that if there is a number  $a$  which is larger than or equal to all numbers  $x \in A$ , then there is a smallest number with this property.*

For instance, the choice  $A = \{x \in \mathbb{Q}, x^2 < 2\}$  gives the supremum  $\sup A$  which is called  $\sqrt{2}$ ; the square root of two.

An immediate consequence of this axiom is the existence of the infima for any non-empty set of real numbers bounded from below. Observe that changing the sign of all the numbers in a set interchanges suprema and infima.

For the formal construction, it is necessary to know whether or not there is such a set  $\mathbb{R}$  with the operations and ordering relation which satisfies the thirteen axioms. So far,

column more functional, we decided to present problems related to such (and other) topological notions a few later, after mastering the notion of limits.

**5.B.3. Convergence and divergence.** Based on the formal definition of the convergence or the divergence of a sequence, prove the following:

- (a)  $a_n = \frac{1}{n} \rightarrow 0$ , as  $n \rightarrow +\infty$ ,
- (b)  $a_n = \frac{1}{n^2} \rightarrow 0$ , as  $n \rightarrow +\infty$ ,
- (c)  $a_n = \frac{n}{n+1} \rightarrow 1$ , as  $n \rightarrow +\infty$ ,
- (d) if  $0 < x < 1$  then  $a_n = x^n \rightarrow 0$ , as  $n \rightarrow +\infty$ ,
- (e) if  $x > 1$  then  $a_n = x^n \rightarrow +\infty$ , as  $n \rightarrow +\infty$ ,
- (f)  $a_n = n^2 \rightarrow +\infty$ , as  $n \rightarrow +\infty$ .

**Solution.** (a) Let  $\varepsilon > 0$  be given. We have to find some positive integer  $N$  such that  $|\frac{1}{n} - 0| < \varepsilon$  for all  $n \geq N$ . Clearly, the condition  $|\frac{1}{n} - 0| < \varepsilon$  is equivalent to  $\frac{1}{n} < \varepsilon$ , i.e.,  $n > \frac{1}{\varepsilon}$ . Hence taking  $N \in \mathbb{Z}_+$  with  $N > \frac{1}{\varepsilon}$  we see that  $\frac{1}{n} \leq \frac{1}{N} < \varepsilon$  so that  $|\frac{1}{n} - 0| < \varepsilon$  for all  $n \geq N$ . This means that  $\frac{1}{n} \rightarrow 0$ .

(b) For any  $\varepsilon > 0$  we see that  $\frac{1}{n^2} < \varepsilon$  is equivalent to  $n^2 > \frac{1}{\varepsilon}$ , that is,  $n > \frac{1}{\sqrt{\varepsilon}}$ . Thus, taking  $N \in \mathbb{Z}_+$  with  $N \geq 1/\sqrt{\varepsilon}$  we have  $\frac{1}{n^2} < \varepsilon$ , for all  $n \geq N$ . It follows that  $\frac{1}{n^2} \rightarrow 0$ .

(c) Again, for given  $\varepsilon > 0$  we wish to find some  $N \in \mathbb{Z}_+$  such that  $|\frac{n}{n+1} - 1| < \varepsilon$ , provided that  $n \geq N$ . Notice that the condition  $|\frac{n}{n+1} - 1| < \varepsilon$  is equivalent to  $\frac{1}{n+1} < \varepsilon$ , i.e.,  $\frac{1}{\varepsilon} < n + 1$ . Hence, choosing  $N$  with  $N > \frac{1}{\varepsilon} - 1$  we see that the inequality  $n \geq N$  implies  $n + 1 > \frac{1}{\varepsilon}$ , and hence  $|\frac{n}{n+1} - 1| < \varepsilon$ . Thus  $\frac{n}{n+1} \rightarrow 1$ .

(d) We have  $0 < x < 1$  and thus  $\frac{1}{x} > 1$ . Hence  $\frac{1}{x} = 1 + h$ , for some  $h > 0$ , and by applying the binomial rule  $\sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$  for the power  $(\frac{1}{x})^n$  we get

$$\frac{1}{x^n} = (1 + h)^n = (1 + nh + \dots + h^n) > 1 + nh.$$

Thus  $0 < x^n < \frac{1}{1+nh}$ . Since  $\frac{1}{1+nh} \rightarrow 0$  we will also have  $x^n \rightarrow 0$ , as  $n \rightarrow \infty$ . For instance, given some  $\varepsilon > 0$  we see that  $0 < x^n < \varepsilon$  is true whenever  $\frac{1}{1+nh} < \varepsilon$ . Hence for  $n \geq N$  with  $N$  being the least positive integer larger than  $\frac{(1/\varepsilon)-1}{h}$ , we are done.

(e) Let us describe a proof that again is based on the binomial theorem. Since  $x > 1$  we may write  $x = 1 + h$  for some  $h > 0$  and by the binomial rule we deduce that  $x^n \geq nh$  for all  $n$ . Since  $nh \rightarrow \infty$  it follows that  $x^n \rightarrow \infty$  as well. In fact in this case we say that the sequence  $(x^n)$  “diverges to infinity”. It is useful to state this as a general rule: For a sequence  $(a_n)$  saying that  $a_n \rightarrow +\infty$  we essentially mean that for any given  $\zeta > 0$  there exists some  $N \in \mathbb{Z}_+$  such that  $a_n > \zeta$  for all  $n \geq N$ . Similarly, saying that a sequence  $(a_n)$  diverges to  $-\infty$  we essentially mean that for any given  $\zeta > 0$  there exists  $N \in \mathbb{Z}_+$  such that  $a_n < -\zeta$  for all  $n \geq N$ .

Let us finally consider case (f). Given  $\zeta > 0$ , the condition

only the rational numbers have been constructed formally. These form an ordered field. This means that  $\mathbb{Q}$  satisfies the axioms (R1) – (R12), and this can easily be verified.

We do not go into details here of the consistent construction of the real numbers now. We will be satisfied with an intuitive idea of the real line, and we will work with the axioms (R1) through R(13). But we shall come back to this issue in a more general framework in chapter 7, see the paragraph 5.2.4 and the discussion started in 7.3.6. Actually, we shall see, that if the real numbers can be constructed, then the construction is unique up to isomorphism. This is a bijection preserving all algebraic structures of two different realizations of the field  $\mathbb{R}$ .

**5.2.2. The complex plane.** Recall that the complex numbers are given as pairs of real numbers. We usually write them as  $z = \text{Re } z + i \text{Im } z$ . Therefore, the plane  $\mathbb{C} = \mathbb{R}^2$  is an appropriate image of the complex numbers.

With addition and multiplication, the complex numbers satisfy the axioms (R1)-(R9) and thus form a field. There is, however, no natural ordering defined on them which would satisfy the axioms (R10)-R(13). Nevertheless, we work with them, since extending real scalars to the complex numbers is highly advantageous or sometimes even necessary.

There is an important operation on the complex numbers called *complex conjugation*. It is the reflection symmetry with respect to the line of real numbers. We denote it by a bar over the number  $z \in \mathbb{C}$ :

$$\bar{z} = \text{Re } z - i \text{Im } z.$$

It changes the sign of the imaginary part. Since for  $z = x + iy$ ,

$$z \cdot \bar{z} = (x + iy)(x - iy) = x^2 + y^2,$$

this value expresses the squared distance of the complex numbers from the origin. The square root of this non-negative real number is called the absolute value of the complex number  $z$ ; written

$$(1) \quad |z|^2 = z \cdot \bar{z}.$$

The absolute value can be defined on any ordered field of scalars  $\mathbb{K}$ . Define the *absolute value*  $|a|$  as follows:

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0. \end{cases}$$

For any numbers  $a, b \in \mathbb{K}$ ,

$$(2) \quad |a + b| \leq |a| + |b|.$$

This property is called the *triangle inequality*. It holds also for the absolute value of the complex numbers.

For the fields of rational numbers or real numbers, both of which are subfields of the complex numbers, both definitions of the absolute value coincide. The absolute value must be understood in the context of which ever field  $\mathbb{K}$  of rational, real, or complex numbers is involved. The triangle inequality holds in all these cases.



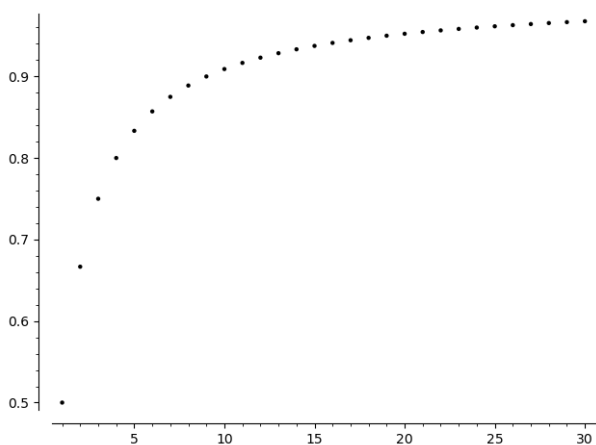
$n^2 > \zeta$  is satisfied if  $n > \sqrt{\zeta}$ . Thus, for  $N > \sqrt{\zeta}$  we see that  $n^2 > \zeta$  for all  $n \geq N$ . This implies that  $n^2 \rightarrow +\infty$ .  $\square$

**5.B.4.** Use Sage to plot the first 30 terms of the sequence  $(a_n)$  given in Problem 5.B.3, case (c).

**Solution.** Sage provides many alternatives to plot a sequence  $(a_n)$ . Here we utilized a method based on the object `Graphics()`, which is employed when initializing a for loop of various graphics objects (further techniques will be analyzed in the sequel). Therefore, to plot the first 30 terms of the sequence  $(\frac{n}{n+1})_{n \in \mathbb{Z}_+}$ , type

```
p=Graphics()
for n in xrange(1, 30+1):
    p=p+points((n, n/(n+1)), color="black")
show(p)
```

This produces the figure given here



For many of the solutions presented in 5.B.3, we have repeatedly used the fact that for every real number  $x$  there exists a natural number  $n \in \mathbb{N}^*$  such that  $n > x$ . This is the so-called Archimedean property of  $\mathbb{R}$ , and it will be used also below without mentioning it explicitly. In fact, it is not hard to prove that the Archimedean property is equivalent to say that for each positive real  $x$  there exists  $n \in \mathbb{N}^*$  such that  $\frac{1}{n} < x$ .



**5.B.5. Infinite limits.** Show that if  $a_n \rightarrow +\infty$  and  $b_n \rightarrow +\infty$ , then  $a_n \cdot b_n \rightarrow +\infty$ . Also, assuming  $a_n \rightarrow a > 0$  and  $b_n \rightarrow +\infty$ , show that  $a_n \cdot b_n \rightarrow +\infty$ . What is the situation for  $a < 0$  and for  $a = 0$ ?

**Solution.** (Hints) First, note that in this context the rule of product of limits must be carefully extended (see 5.2.13 for the product rule). This is because at least one of the limits in the given statements is infinite, refer to the discussion in 5.2.14 for further details. To prove the statement we will apply the definition given in case (e) in 5.B.3. We will describe the proof for the second statement only, and the proof of the first is left for practice.

**5.2.3. Convergence of a sequence.** We wish to formalize the notion of a sequence of numbers approaching a limit. The key object of interest is a sequence of numbers  $a_i$ , where the index  $i$  usually goes through all the natural numbers. Denote the sequences loosely either as  $a_0, a_1, \dots$ , or as infinite vectors  $(a_0, a_1, \dots)$ , or as  $(a_i)_{i=1}^\infty$ .



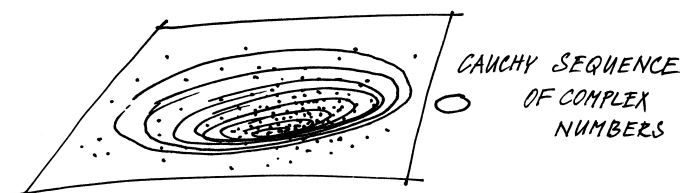
CAUCHY<sup>5</sup> SEQUENCES

Consider a sequence  $(a_0, a_1, \dots)$  of elements of  $\mathbb{K}$  such that for any fixed positive number  $\varepsilon > 0$ ,

$$|a_i - a_j| < \varepsilon.$$

for all but finitely many terms  $a_i, a_j$  of the sequence.

In other words, for any fixed  $\varepsilon > 0$ , there is an index  $N$  such that the above inequality holds for all  $i, j > N$ . Loosely put, the elements of the sequence are eventually arbitrarily close to each other. Such a sequence is called a *Cauchy sequence*.



Intuitively, either all but finitely many of the sequence's terms are equal, (then  $|a_i - a_j| = 0$  from some index  $N$  on), or they "approach" some particular value. This is easily imaginable in the complex plane. Choose an arbitrarily small disc (with radius equal to  $\varepsilon$ ). Suppose a Cauchy sequence is given. It must be possible to put the disc into the complex plane in such a way that it covers all but finitely many of the elements of the infinite sequence  $a_i$ . Imagine that such discs have very small radii, and all contain a number  $a$ ; see the diagram. If such a value  $a \in \mathbb{K}$  exists for a Cauchy sequence, we would expect the sequence to have the property of *convergence*:

CONVERGENT SEQUENCES

A sequence  $(a_i)_{i=0}^\infty$  converges to a value  $a$  if and only if for any positive real number  $\varepsilon$ ,

$$|a_i - a| < \varepsilon$$

for all but finitely many indices  $i$ . Notice that the set of those  $i$  for which the inequality does not hold may depend on  $\varepsilon$ . The number  $a$  is called the *limit* of the sequence  $(a_i)_{i=0}^\infty$ , we write

$$\lim_{i \rightarrow \infty} a_i = a.$$

If a sequence  $a_i \in \mathbb{K}, i = 0, 1, \dots$ , converges to  $a \in \mathbb{K}$ , then for any fixed positive  $\varepsilon, |a_i - a| < \varepsilon$  for all  $i$  greater than

<sup>5</sup>Augustin-Louis Cauchy (1789-1857) was a French mathematician pioneering a rigorous approach to infinitesimal analysis. He was very productive, wrote about 800 research articles. There are dozens of concepts and theorems named after him.

By assumption,  $\lim_{n \rightarrow +\infty} a_n = a > 0$ , so choosing  $\varepsilon = a/2 > 0$  we can find  $N \in \mathbb{Z}_+$  such that  $|a_n - a| < \frac{a}{2}$  for all  $n \geq N$ . This implies that  $\frac{a}{2} < a_n < \frac{3a}{2}$ , for all  $n \geq N$ , and so  $a_n > \frac{a}{2}$  for  $n \geq N$ . Also, choosing any  $\zeta > 0$  there exists  $M \in \mathbb{Z}_+$  such that  $b_n > \zeta$  for all  $n \geq M$ . Hence, for  $n \geq L := \max\{M, N\}$  both conditions hold and in particular  $a_n \cdot b_n > \frac{a}{2} \cdot \zeta > 0$ . Taking  $\zeta = \frac{2\tilde{\zeta}}{a}$  for some  $\tilde{\zeta} > 0$  we get  $a_n \cdot b_n > \tilde{\zeta}$  for all  $n \geq L$ . Thus  $a_n \cdot b_n \rightarrow +\infty$ , and  $a \cdot (+\infty) = +\infty$  for  $a > 0$ . For  $a < 0$  we have  $a_n \cdot b_n \rightarrow -\infty$ , and thus  $a \cdot (+\infty) = -\infty$ , for  $a < 0$ . However, the case  $a = 0$  will produce a so-called *indeterminate form*, namely  $0 \cdot (+\infty)$ , which is excluded.  $\square$

**5.B.6.** Using the results from 5.B.3 explain why:

- (a) The sequence  $(a_n = \frac{1}{n} + 3^n)$  diverges to  $+\infty$ .
- (b) The sequence  $(b_n = n^2 + 2^n)$  diverges to  $+\infty$ .
- (c) The sequence  $(c_n = n^2 \cdot 3^n)$  diverges to  $+\infty$ .
- (d) The sequence  $(d_n = \frac{n \cdot 4^n}{n+1})$  diverges to  $+\infty$ .  $\circ$

- 5.B.7.** (a) Prove that  $(a_n = \frac{1}{2^n})$  tends to zero as  $n \rightarrow \infty$ .  
 (b) Prove that  $(b_n = (-1)^{n+1})$  is divergent.  
 (c) Prove that  $(c_n = \sin(n\pi/2))$  is divergent.  
 (d) Find sequences  $(x_n)$  and  $(y_n)$  with infinite limits, such that  $\lim_{n \rightarrow \infty} (x_n + y_n) = 1$  and  $\lim_{n \rightarrow \infty} (x_n y_n^2) = +\infty$ .  $\circ$

**5.B.8.** Show that  $\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$  (notice  $\sqrt[n]{n} = n^{\frac{1}{n}}$ ). Then present a proof via Sage.

**Solution.** Apparently, for all naturals  $n \geq 1$  we have  $\sqrt[n]{n} \geq 1$ . So we can set  $\sqrt[n]{n} = 1 + a_n$  for certain numbers  $a_n \geq 0$ . Now, by the binomial theorem we see that

$$n = (1 + a_n)^n = 1 + \binom{n}{1} a_n + \binom{n}{2} a_n^2 + \dots + a_n^n,$$

for all natural numbers  $n \geq 2$ . Hence we have the bound

$$n \geq \binom{n}{2} a_n^2 = \frac{n(n-1)}{2} a_n^2,$$

for all naturals  $n \geq 2$ , which leads to  $0 \leq a_n \leq \sqrt{\frac{2}{n-1}}$  for all such  $n$ . Having established these inequalities, we can now use the squeeze theorem (see 5.2.12). This implies that  $\lim_{n \rightarrow \infty} a_n = 0$ , and hence  $\lim_{n \rightarrow \infty} \sqrt[n]{n} = \lim_{n \rightarrow \infty} (1 + a_n) = 1$ . One should finally mention that it is possible to prove the result by means of the so called ‘l’Hopital’s rule’, introduced in 5.3.10. However, such a procedure is based on the notion of derivatives and hence we will return to this later, see 5.E.109.

For computing limits in Sage we use the command `limit` or its alias `lim`. We also recall that the mathematical symbol  $\infty$  is represented either by `oo` or by `Infinity` (or `infinity`). Let us now illustrate the implementation of these rules via our example. So, an appropriate cell has the form

```
n=var("n"); a(n)=n**(1/n); lim(a(n), n=oo)
```

Or, as we said, one could replace the last command by `lim(a(n), n = Infinity)`. Notice that first one should introduce the variable  $n$ , as a symbolic variable. This is because

a certain  $N \in \mathbb{N}$ . By the triangle inequality,

$$|a_i - a_j| = |a_i - a + a - a_j| < |a_i - a| + |a - a_j| < 2\varepsilon,$$

for all pairs of indices  $i, j \geq N$ . Thus:

**Lemma.** Every convergent sequence is a Cauchy sequence.

However, in the field of rational numbers, it can happen that for a Cauchy sequence a corresponding value  $a$  does not exist. For instance, the number  $\sqrt{2}$  can be approached by a sequence of rational numbers  $a_i$ , thereby obtaining a sequence converging to  $\sqrt{2}$ , but the limit is not rational.

Ordered fields of scalars in which every Cauchy sequence converges are called *complete*. The following theorem states that the axiom (R13) guarantees that the real numbers are such a field:

**Theorem.** Every Cauchy sequence of real numbers  $a_i$  converges to a real number  $a \in \mathbb{R}$ .

**PROOF.** The terms of any Cauchy sequence form a bounded set since any choice of  $\varepsilon$  bounds all but finitely many of them. Let  $B$  be the set of those real numbers  $x$  for which  $x < a_j$  for all but finitely many terms  $a_j$  of the sequence.  $B$  has an upper bound, and thus  $B$  has a supremum as well, by (R13).

Define  $a = \sup B$ . Fix  $\varepsilon > 0$ , and choose  $N$  so that  $|a_i - a_j| < \varepsilon$  for all  $i, j \geq N$ . Then  $a_j > a_N - \varepsilon$  and  $a_j < a_N + \varepsilon$  for all indices  $j > N$ , and so  $a_N - \varepsilon$  belongs to  $B$ , while  $a_N + \varepsilon$  does not. Hence  $|a - a_N| \leq \varepsilon$ , and thus

$$|a - a_j| \leq |a - a_N| + |a_N - a_j| \leq 2\varepsilon$$

for all  $j > N$ . So  $a$  is the limit of the given sequence.  $\square$

**Corollary.** Every Cauchy sequence of complex numbers  $z_i$  converges to a complex number  $z$ .

**PROOF.** Write  $z_i = a_i + i b_i$ . Since  $|a_i - a_j|^2 \leq |z_i - z_j|^2$  and similarly for the values  $b_i$ , both sequences of real numbers  $a_i$  and  $b_i$  are Cauchy sequences. They converge to  $a$  and  $b$ , respectively. It is easily verified that  $z = a + i b$  is the limit of the sequence  $z_i$ .  $\square$

**5.2.4. Remark.** The previous discussion proposes a construction method for the real numbers. Proceed similarly to building the integers from the natural numbers (adding in all additive inverses). Build the rational numbers from the integers (adding all multiplicative inverses of non-zero numbers). Then ‘complete’ the rational numbers by adding in all limits of Cauchy sequences.

Cauchy sequences  $(a_i)_{i=0}^\infty$  and  $(b_i)_{i=0}^\infty$  of rational numbers are equivalent if and only if the distances  $|a_i - b_i|$  converge to zero. This is the same as the condition that merging these sequences into a single sequence also yields a Cauchy sequence. For example, a sequence can be formed by selecting alternately terms from the first sequence and the second sequence. Check the properties of the equivalence relations. Clearly the relation is reflexive, it is symmetric (since the distance of the rational numbers is symmetric in its arguments)

our goal is to view  $a(n)$  as a function of  $n$ . Execute yourself the given syntax to practice with Sage.  $\square$

**5.B.9. Limits of sequences.** Compute the following limits, and next verify your answers via Sage:

- (a)  $\lim_{n \rightarrow \infty} \frac{2n^2+3n+1}{n+1}$ , (d)  $\lim_{n \rightarrow \infty} \frac{n+1}{2n^2+3n+1}$ ,  
 (b)  $\lim_{n \rightarrow \infty} \sqrt[n]{1^2 + 2^2 + \dots + n^2}$ , (e)  $\lim_{n \rightarrow \infty} \frac{5^n+1}{7^n+1}$ ,  
 (c)  $\lim_{n \rightarrow \infty} \sqrt{4n^2 + n} - 2n$ , (f)  $\lim_{n \rightarrow \infty} \frac{\sqrt{4n^2+n}}{n}$ .

**Solution.** (a) Set  $a_n = \frac{2n^2+3n+1}{n+1}$ . Multiplying both the numerator and denominator by  $1/n$ , we have

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} \frac{2n + 3 + \frac{1}{n}}{1 + \frac{1}{n}} = \frac{\infty + 3 + 0}{1 + 0} = \infty.$$

To confirm this result in Sage, give the cell

```
n=var("n"); a(n)=(2*n^2+3*n+1)/(n+1)
lim(a(n), n=infinity)
```

Sage’s answer is +Infinity.

(b) Let us present a solution based on a very useful theoretical result, the so called “squeeze theorem” presented in 5.2.12. In particular, set  $a_n = \sqrt[n]{1^2 + 2^2 + \dots + n^2}$ . Then for any positive natural  $n$  we have

$$b_n = \sqrt[n]{n} \leq a_n \leq c_n = \sqrt[n]{\underbrace{n^2 + n^2 + \dots + n^2}_{n\text{-terms}}} = \sqrt[n]{n^3}.$$

By 5.B.8 we know that  $b_n \rightarrow 1$ . Moreover, since  $\sqrt[n]{n^3} = (n^3)^{\frac{1}{n}} = n^{\frac{3}{n}} = (n^{\frac{1}{n}})^3$ , we also conclude that  $c_n \rightarrow 1$ , as  $n \rightarrow +\infty$ . The squeeze theorem now applies and gives that  $a_n \rightarrow 1$ .

In Sage one can verify these claims as follows:

```
n, k=var("n, k")
a(n)=sum(k^2, k, 1, n)**(1/n) # declare (a_n)
f(n)=(n^3)**(1/n); g(n)=n**(1/n)
print(bool(a(n)<=f(n) for n in range(1, 1000)))
print(bool(a(n)>=g(n) for n in range(1, 1000)))
print(lim(a(n), n=oo))
```

Notice in the fourth and fifth line we used the bool command to test the two posed inequalities, for many  $n$ . However, one could directly compute the limit in question (hence, one may avoid to include these two lines, together with the definition of the sequences  $f(n)$  and  $g(n)$ ).

(c) Set  $a_n = \sqrt{4n^2 + n} - 2n$ . Here it applies the following classical trick:

$$\begin{aligned} \lim_{n \rightarrow \infty} a_n &= \lim_{n \rightarrow \infty} \frac{(\sqrt{4n^2 + n} - 2n)(\sqrt{4n^2 + n} + 2n)}{\sqrt{4n^2 + n} + 2n} \\ &= \lim_{n \rightarrow \infty} \frac{n}{\sqrt{4n^2+n}+2n} = \lim_{n \rightarrow \infty} \frac{1}{\frac{\sqrt{4n^2+n}}{n} + 2} = \frac{1}{4}. \end{aligned}$$

(d) We leave this for practice, since can be treated as case (a). Verify yourself that the given limit equals to  $\frac{1}{\infty} = 0$ .

and transitivity follows easily from the triangle inequality. Thus, we may define  $\mathbb{R}$  as the set of equivalence classes on the above set of sequences.

We introduce algebraic structures on this set  $\mathbb{R}$  and check their properties. Of course, the rational numbers can be represented by constant sequences, so that  $\mathbb{Q} \subset \mathbb{R}$ , as expected. Next, define the sum and product of sequences representing them, respectively. It is easy to check that the results represent a class independent of the choices.

Ordering is dealt with similarly. Here it is required to prove that  $a \leq b$  if and only if there are representatives with  $a_i \leq b_i$ . Finally it is necessary to show that all Cauchy sequences in  $\mathbb{R}$  converge. We do not go into details now and advise the reader to return back and check all the details when going through the full discussion of the completion of metric spaces in the paragraph 7.3.6. The arguments used there with the real scalars replaced by rational ones provide an adequate proof. The arguments proving that the axioms (R1)–(R13) define the real numbers uniquely up to isomorphism are also to be found there.

**5.2.5. Closed sets.** For further work with real or complex numbers, we need to understand the notions of closeness, boundedness, convergence, and so on. These concepts belong to the topic “topology”<sup>6</sup>. As before, we work with  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{K} = \mathbb{C}$ . We advise the reader to draw many diagrams for all the concepts and their properties for both the real line and the complex plane.

For any subset  $A$  of points in  $\mathbb{K}$ , we are interested not only in the points belonging to  $a \in A$ , but also in the ones which can be approached by limits of sequences in  $A$ .

LIMIT POINTS OF A SET

Let  $A$  be a subset of  $\mathbb{K}$ . A point  $x \in \mathbb{K}$  is called a *limit point* of  $A$  if and only if there is a sequence  $a_0, a_1, \dots$  of elements in  $A$  such that all its terms differ from  $x$ , yet its limit is  $x$ .

Notice that a limit point of a set may or may not belong to the set.

For every non-empty set  $A \subset \mathbb{K}$  and a fixed point  $x \in \mathbb{K}$ , the set of all distances  $|x - a|, a \in A$ , is a set of real numbers bounded from below, and so it has an infimum  $d(x, A)$ , which is called the *distance of the point  $x$  from the set  $A$* . Notice that  $d(x, A) = 0$  if and only if either  $x \in A$  or if  $x$  is a limit point of  $A$ . (We suggest the reader proves this in detail directly from the definitions.)

<sup>6</sup>The name of this mathematical discipline comes from the Greek “studying the shape” (topos + logos). The main concepts are built on the formalism of open and closed sets, compactness etc. We use the same names here but only in the realm of real and complex numbers. Later on, we go further in metric spaces in chapter 7.

(e) This case can be treated using the conclusions of 5.B.3. Hence we get

$$\lim_{n \rightarrow \infty} \frac{5^n + 1}{7^n + 1} = \lim_{n \rightarrow \infty} \frac{\left(\frac{5}{7}\right)^n + \frac{1}{7^n}}{1 + \frac{1}{7^n}} = \frac{0 + 0}{1 + 0} = 0.$$

In Sage for this result work as above, i.e.,

```
n=var("n"); f(n) = ((-5)^n + 1) / ((7^n) + 1)
lim(f(n), n=oo)
```

(f) Set  $a_n = \frac{\sqrt{4n^2+n}}{n}$ . For any positive natural  $n$  we have that

$$b_n = \frac{\sqrt{4n^2}}{n} < a_n < c_n = \frac{\sqrt{4n^2 + n + \frac{1}{16}}}{n}.$$

Moreover, it is easy to see that  $\lim_{n \rightarrow \infty} b_n = 2 = \lim_{n \rightarrow \infty} c_n$ . Thus, by the squeeze theorem it follows that  $\lim_{n \rightarrow \infty} a_n = 2$ .  $\square$

**5.B.10.** Given a non-empty bounded set  $A \subseteq \mathbb{R}$ , set  $a := \sup A$ . Show that there exists a sequence  $(a_n)$  of  $A$  tending to  $a$  (notice a similar statement holds for  $\inf A$ ).

**Solution.** By assumption  $a$  is the supremum of  $A$ , so  $a$  is an upper bound of  $A$  and for every  $\zeta > 0$  there exists  $x \in A$  such that  $x > a - \zeta$ . To prove this fix some  $\zeta > 0$  and suppose in the contrary that  $x \leq a - \zeta$ . Then  $a - \zeta$  must be also an upper bound of  $A$ , a contradiction since  $a = \sup A$ , so  $a$  is the least upper bound. Thus we can find some  $x \in A$  with  $x > a - \zeta$  (note that the converse is also true). Choosing now  $\zeta = 1/n$  for  $n \in \mathbb{Z}_+$ , so  $\zeta > 0$ , we may denote this element by  $x_n \in A$ , such that  $x_n > a - \frac{1}{n}$ . Since  $x_n \leq a$ , we finally arrive to the inequality  $0 \leq a - x_n < 1/n$ . Fix now some  $\varepsilon > 0$  and choose  $N \in \mathbb{Z}_+$  with  $1/N < \varepsilon$ . Then for all  $n \geq N$  we get  $|a - x_n| < 1/n \leq 1/N < \varepsilon$ , hence  $x_n \rightarrow a$  as  $n \rightarrow \infty$ .  $\square$

A sequence  $(a_n)$  is said to be bounded above (respectively, bounded below), if there exists  $M \in \mathbb{R}$  such that  $a_n \leq M$  (respectively,  $a_n \geq M$ ), for all  $n$  that  $(a_n)$  is defined. A sequence that is bounded above and below is called bounded.

Obviously, a sequence that diverges to  $+\infty$  (respectively,  $-\infty$ ) is not bounded above (respectively, is not bounded below). A sequence  $(a_n)$  is said to be increasing (respectively, decreasing) if  $a_n \leq a_{n+1}$  (respectively,  $a_n \geq a_{n+1}$ ) for all  $n$  that  $(a_n)$  is defined. If we have  $a_n < a_{n+1}$  (respectively,  $a_n > a_{n+1}$ ) for all  $n$ , then the sequence  $(a_n)$  is called strictly increasing (respectively, strictly decreasing). For instance, the sequence  $(n^2)$  is strictly increasing, while the sequence  $(1/n)$  is strictly decreasing.

**5.B.11. Monotone sequence theorem.** Show that every bounded monotonic sequence is convergent.

**Solution.** This extremely useful result can be equivalently rephrased as follows: “Every increasing sequence  $(a_n)$  which is bounded above converges to  $\sup\{a_n : n \in \mathbb{N}^*\}$ , and similarly every decreasing sequence  $(b_n)$  that is bounded below

CLOSED SETS

The closure  $\bar{A}$  of a set  $A \subset \mathbb{K}$  is the set of those points which have zero distance from  $A$  (note that the distance from the empty set of points is undefined, therefore  $\bar{\emptyset} = \emptyset$ ).

A closed subset in  $\mathbb{K}$  is a set which coincides with its closure. A set is closed if it contains all of its limit points.

On the real line, a closed interval

$$[a, b] = \{x \in \mathbb{R}, a \leq x \leq b\}$$

of real numbers, where  $a$  and  $b$  are fixed real numbers is a closed set.

The sets  $(-\infty, b]$ ,  $[a, \infty)$ , and  $(-\infty, \infty)$  are also closed sets.

A closed set may also be formed by a sequence of real numbers without a limit point, or a sequence with a finite number of limit points together with these points.

The unit disc (including its boundary circle) in the complex plane is another example of a closed set.

An arbitrary intersection of closed sets is again a closed set. A finite union of closed sets is again a closed set. Indeed, if all of the points of some sequence belong to the considered intersection of closed sets, then they belong to each of the sets, and so do all the limit points. However, if we wanted to say the same about an arbitrary union, we would get in trouble: singleton sets are closed, but a sequence of points created from them may not be. On the other hand, if we restrict our attention to finite unions and consider a limit point of some converging sequence lying in this union, then the limit point must also be the limit point of any subsequence, especially the one lying in only one of the united sets. As this set is assumed to be closed, the limit point lies in it, and thus it lies in the union.

**5.2.6. Open sets.** There is another useful type of subset of the real numbers: open intervals

$$(a, b) = \{x \in \mathbb{R}; a < x < b\},$$

where, again,  $a$  and  $b$  are fixed real numbers or infinite values  $\pm\infty$ . It is an open set, in the following sense:

OPEN SETS AND NEIGHBOURHOODS OF POINTS

An open set in  $\mathbb{K}$  is a set whose complement is a closed set.

A neighbourhood of a point  $a \in \mathbb{K}$  is an open set  $\mathcal{O}$  which contains  $a$ . If the neighbourhood is defined as

$$\mathcal{O}_\delta(a) = \{x \in \mathbb{K}, |x - a| < \delta\}$$

for some positive number  $\delta$ , then we call it the  $\delta$ -neighbourhood of the point  $a$ .

For real numbers, the  $\delta$ -neighbourhood of a point  $a$  is the open interval of length  $2\delta$ , centered at  $a$ . In the complex plane, it is the disc of radius  $\delta$ , also centered at  $a$ .

Notice that for any set  $A$ ,  $a \in \mathbb{K}$  is a limit point of  $A$  if and only if every neighbourhood of  $a$  contains at least one point  $b \in A$ ,  $b \neq a$ .

**Lemma.** A set  $A \subset \mathbb{K}$  of numbers is open if and only if every point  $a \in A$ , has a neighbourhood contained in  $A$ .

converges to  $\inf\{b_n : n \in \mathbb{N}^*\}$ . We will prove the first statement and an analogous method applies for the second one. Let  $(a_n)$  be a sequence and assume for simplicity that  $n \in \mathbb{N}^*$ . Assume also that  $(a_n)$  is bounded above with  $a_n \leq a_{n+1}$  for all  $n \in \mathbb{N}^*$ . Since  $(a_n)$  is bounded above, its range  $\{a_n : n \in \mathbb{N}^*\}$  is also bounded above (as a set), and hence  $\sup\{a_n : n \in \mathbb{N}^*\}$  exists. Set  $a := \sup\{a_n : n \in \mathbb{N}^*\}$ . Recall that if  $a$  is the supremum of a non-empty set  $A$ , then  $a$  is an upper bound of  $A$  and for every  $\zeta > 0$  there exists  $x \in A$  such that  $x > a - \zeta$  (see the proof of 5.B.10). In our case this means that for  $\zeta > 0$  be given, it exists a natural  $N$  with  $a - \zeta < a_N$ , that is,  $a - a_N < \zeta$ . However,  $(a_n)$  is increasing, so for all  $n \geq N$  we have  $a_n \geq a_N$  and in particular  $a - \zeta < a_N \leq a_n \leq a < a + \zeta$ , for all such  $n$ . Thus  $|a_n - a| < \zeta$  for all  $n \geq N$ , i.e.,  $a_n \rightarrow a$ .  $\square$

**5.B.12.** Present a “counterexample” verifying that the converse of the monotone sequence theorem fails. Moreover, give an example of a monotone sequence which is not convergent.  $\circ$

**5.B.13.** Given the sequences  $(a_n = \frac{\sin(n)}{n})$ ,  $(b_n = n^2 + 1)$  and  $(c_n = \frac{(-1)^{n+n}}{n})$ , with  $n \in \mathbb{Z}_+$  for all three cases, locate them which are bounded.  $\circ$

**5.B.14.** Indicate the convergent sequences between those given in the previous problem 5.B.13, and determine their limits. Repeat for the sequences  $f_n = \frac{1}{\sqrt{n}}$  and  $g_n = \frac{n!}{n^n}$ , where  $n \in \mathbb{Z}_+$ .  $\circ$

**5.B.15.** Let  $a > 0$ . Show that  $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$ .

**Solution.** Once more, squeeze theorem provides a traditional way to treat this limit, as in 5.B.8. Let us however discuss a different method. First notice that we may assume that  $0 < a < 1$ , since otherwise the argument applies for  $1/a$ . For such  $a$  it is easy to see that given sequence  $(a_n = \sqrt[n]{a})_{n \in \mathbb{Z}_+}$  is monotonically increasing. Moreover, it is bounded above by 1. A proof of this fact is left to the reader, but one may like to illustrate the situation in Sage, either via the cell

```
n=var("n"); a=0.2; a(n)=a**(1/n)
bool(a(n)<1 for n in range(1, 5000))
```

which returns True, or by plotting many terms of  $(a_n)$  (continue typing in the previous cell)<sup>4</sup>

```
list_plot([a(n) for n in range(1, 100)])
```

Thus  $(a_n)$  converges, and we may assume that  $\lim_{n \rightarrow +\infty} a_n = \ell \in \mathbb{R}$ .

Assume that  $\ell < 1$ . Using the inequality  $\sqrt[n]{a} \leq \ell$  we get  $a \leq \ell^n$ , with  $\ell^n \rightarrow 0$  by 5.B.3. This gives a contradiction, since  $a > 0$ . Taking into account that  $(a_n)$  is bounded above by 1, we deduce that  $\ell = 1$ . For a Sage verification one can continue in the very first cell used above, by adding the following command: `lim(a(n), n = oo)`.  $\square$

<sup>4</sup>Notice this gives an alternative to plot sequences via Sage.

**PROOF.** Let  $A$  be an open set and let  $a \in A$ . If there is no neighbourhood of the point  $a$  inside  $A$ , then there is a sequence  $a_n \notin A$ ,  $|a - a_n| \leq 1/n$ . But then the point  $a \in A$  is a limit point of the set  $\mathbb{K} \setminus A$ , which is impossible since the complement of  $A$  is closed.

Suppose every  $a \in A$  has a neighbourhood contained in  $A$ . This prevents a limit point  $b$  of the set  $\mathbb{K} \setminus A$  to lie in  $A$ . Thus the set  $\mathbb{K} \setminus A$  is closed, and so  $A$  is open.  $\square$

From this lemma, it follows immediately that any union of open sets is an open set. A finite intersection of open sets is also an open set.

**5.2.7. Bounded and compact sets.** The closed and open sets are basic concepts of *topology*. Without going into deeper connections, the above material describes the *topology of the real line* and the *topology of the complex plane*. The following concepts are extremely useful:



#### BOUNDED AND COMPACT SETS

A set  $A$  of rational, real, or complex numbers is called *bounded* if and only if there is a positive real number  $r$  such that  $|z| \leq r$  for all numbers  $z \in A$ . Otherwise, the set is called *unbounded*.

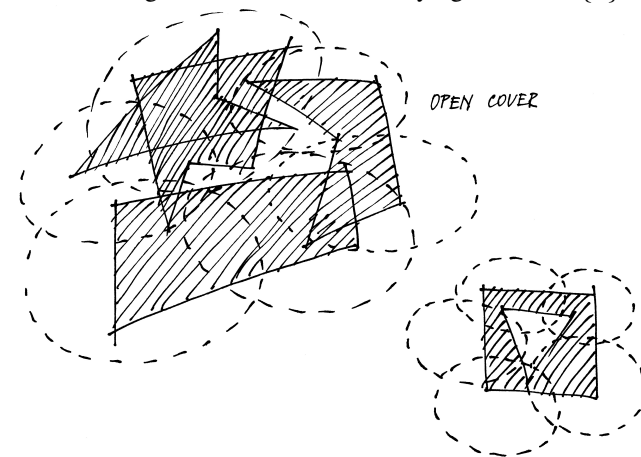
A set which is both bounded and closed is called *compact*.

An *interior point* of a set  $A$  is a point such that one of its neighbourhoods is contained in  $A$ .

A *boundary point* of a set  $A$  is a point for which all its neighbourhoods have a non-trivial intersection with both  $A$  and its complement  $\mathbb{K} \setminus A$ . A boundary point of the set  $A$  may or may not belong to it.

An *open cover* of a set  $A$  is such a collection of open sets  $U_i, i \in I$ , that its union contains the whole of  $A$ .

An *isolated point* of a set  $A$  is a point  $a \in A$  such that there is a neighbourhood  $N$  of  $a$  satisfying  $N \cap A = \{a\}$ .



**5.2.8. Theorem.** All subsets  $A \subset \mathbb{K}$  of real or complex numbers satisfy:

- (1) a non-empty set  $A \subset \mathbb{R}$  is open if and only if it is a union of countably (or finitely) many open intervals; similarly



Often we are interested in the limit of a sequence  $s_n$ , where the terms are partial sums of some other given sequence  $a_n$ . Let us describe such an example and more applications are described in Section D, which is about infinite series.

**5.B.16. Geometric series.** Find the limit of the sequence with general term

$$s_n = 1 + x + x^2 + \cdots + x^n, \quad n = 0, 1, 2, \dots,$$

for  $x$  with  $|x| < 1$ .

**Solution.** First observe that

$$\begin{aligned} s_n(1-x) &= (1+x+\cdots+x^n) - x(1+x+\cdots+x^n) \\ &= 1+x+\cdots+x^n - x - \cdots - x^n - x^{n+1} \\ &= 1 - x^{n+1}. \end{aligned}$$

So for  $x \neq 1$  we have  $s_n = \frac{1-x^{n+1}}{1-x}$ , with  $n \in \mathbb{N}$ . Recall now by 5.B.3 that  $\lim_{n \rightarrow \infty} x^n = 0$ , provided that  $|x| < 1$ . Thus for such  $x$  we get

$$\lim_{n \rightarrow \infty} s_n = \lim_{n \rightarrow \infty} \frac{1-x^{n+1}}{1-x} = \frac{1}{1-x},$$

and it follows that

$$\sum_{n=0}^{\infty} x^n = \lim_{n \rightarrow \infty} s_n = \frac{1}{1-x}, \quad |x| < 1.$$

This is the so called *geometric series* and for those  $x$  satisfying  $|x| < 1$ , the geometric series is a finite number. For instance

$$\sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n - 1 = \frac{1}{1-\frac{1}{2}} - 1 = 2 - 1 = 1,$$

that is,

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots = 1.$$

Very often Sage allows us to compute such sums, by applying for example the following cell (which indeed returns  $\frac{-1}{x-1}$ ):

```
x, n=var("x, n"); assume(abs(x) < 1)
print(sum(x^n, n, 0, oo))
```

This is preliminary example of the `sum` method in Sage, which we will analyze more in Section D. On the other hand, the geometric series will appear in the foreground very often, hence we encourage you to devote some additional time with it (e.g., compute in Sage some of its partial sums).  $\square$

**5.B.17.** Consider the sequence  $(a_n)$  with

$$a_n = \frac{8}{10} + \frac{8}{10^2} + \cdots + \frac{8}{10^n}, \quad n = 1, 2, \dots$$

that is,  $a_n = 0.88 \cdots 8$ , with 8 repeated  $n$  times. Show that  $a_n$  tends to  $\frac{8}{9}$  as  $n$  tends to  $\infty$ , and then give a proof via Sage.

$A \subset \mathbb{C}$  is open if and only if it is a union of countably (finite) many open discs.

- (2) every point  $a \in A$  is either an interior or a boundary point,
- (3) every boundary point of  $A$  is either an isolated point or a limit point of  $A$ ,
- (4)  $A$  is compact if and only if every infinite sequence contained in  $A$  has a subsequence converging to a point in  $A$ .<sup>7</sup>
- (5)  $A$  is compact if and only if each of its open covers contains a finite subcover of  $A$ .

**PROOF.** (1) Every open set is a union of some neighbourhoods of its points, i.e., we may consider open intervals in reals, or open discs in  $\mathbb{C}$ . So the question that remains is whether it suffices to take countably many of them. Let us first prove the claim for the complex plane. For each  $z \in A$ , there is an open disc  $\mathcal{O}_\delta(z)$  contained in  $A$ , with some  $\delta > 0$ , and let  $\delta_z$  be the supremum of the values of such  $\delta$ . Clearly,  $A = \cup_{z \in A} \mathcal{O}_{\delta_z}(z)$ . Consider an arbitrary  $z \in A$  and pick up  $w$  with both real and imaginary parts rational, such that  $|w - z| < \delta_z/4$ . Thus,  $z \in \mathcal{O}_{\delta_w}(w)$  (draw a picture!) and we have checked that actually  $A$  is the union of the countably many open discs  $\mathcal{O}_{\delta_w}(w)$  for all  $w \in A$  with rational real and imaginary coordinates.

If  $A$  is an open subset in  $\mathbb{R}$ , then we may repeat the above argument with the discs  $\mathcal{O}_\delta(z)$  replaced by the intervals  $\mathcal{O}_\delta(x)$  and  $x \in A$ . Think about the details!

(2) It follows immediately from the definitions that no point can be both an interior and boundary point. Let  $a \in A$  be a point that is not interior. Then there is a sequence of points  $a_i \notin A$  with  $a$  as its limit point. At the same time,  $a$  belongs to each of its neighbourhoods. Thus  $a$  is a boundary point.

(3) Suppose that  $a \in A$  is a boundary point but not isolated. Then, similarly to the reasoning from the previous paragraph, there are points  $a_i$ , this time inside  $A$ , whose limit point is  $a$ .

(4) Suppose  $A \subset \mathbb{R}$  is a compact set, i.e., both closed and bounded. Consider an infinite sequence of points  $a_i \in A$ .  $A$  has both a supremum  $b$  and an infimum  $a$ . Divide the interval  $[a, b]$  into halves:  $[a, \frac{1}{2}(b-a)]$  and  $[\frac{1}{2}(b-a), b]$ . At least one of them contains infinitely many of the terms  $a_i$ . Select this half and one of the terms contained in it; then cut the selected interval into halves. Again, select such a half which contains infinitely many of the sequence's terms and select one



<sup>7</sup>This result for real numbers is usually referred to as the Bolzano-Weierstrass theorem. Karl Weierstrass was a famous German mathematician (1815-1897) and his name is linked to many theorems in Mathematics. Bernard Bolzano (1781-1848) was a Bohemian mathematician, logician, philosopher, theologian and Catholic priest working in Prague at the beginning of the 19th century. He laid the basis of rigorous mathematical analysis a few decades before all the theory was worked out by Weierstrass and others. In particular he was skeptical about the effective use of Leibniz's infinitesimals without the necessary rigour.

**Solution.** We see that

$$\begin{aligned} a_n &= \frac{8}{10} \left( 1 + \frac{1}{10} + \cdots + \frac{1}{10^{n-1}} \right) = \frac{8}{10} \left( \frac{1 - \frac{1}{10^n}}{1 - \frac{1}{10}} \right) \\ &= \frac{8}{9} \left( 1 - \frac{1}{10^n} \right) = \frac{8}{9} - \frac{8}{9} \cdot \frac{1}{10^n}. \end{aligned}$$


This gives  $|a_n - \frac{8}{9}| = \frac{8}{9} \cdot \frac{1}{10^n}$  and hence it is easy to see that the condition  $|a_n - \frac{8}{9}| < \varepsilon$  is equivalent to the inequality  $10^n > \frac{8}{9\varepsilon}$ . Therefore, if  $N$  is a natural satisfying  $10^N > \frac{8}{9\varepsilon}$ , then we have  $|a_n - \frac{8}{9}| < \varepsilon$ , for all  $n \geq N$ . Thus we are done.

An alternative relies on the fact  $\frac{1}{10^n} \rightarrow 0$  as  $n \rightarrow \infty$ . To prove this, notice that for given  $\varepsilon > 0$  and for all  $n \geq N := \log_{10}(1/\varepsilon)$  one has  $|1 - \frac{1}{10^n}| = \frac{1}{10^n} \leq \frac{1}{10^N} = \frac{1}{1/\varepsilon} = \varepsilon$ .

Finally, the following block confirms the result by a less obvious technique (explain why this is true).

```
var ("n, k") ; a=sum(8/(10)^n, n, 1, oo)
bool(a==lim(sum(8/(10)^k, k, 1, n), n=oo))
```


□

 The binomial theorem often combines with the monotone sequence theorem and together provide elegant techniques for studying the convergence of sequences. Such an example encodes our next task, which is about a famous number, the so called “Euler number” denoted by  $e$ , that is, the base of the natural logarithm, see 5.4.1. This satisfies  $2 < e < 3$ , approximately equals to  $e \approx 2.7183$ , and is the limit of the sequence  $(e_n)_{n=1}^\infty$  featured in 5.B.18.<sup>5</sup>

**5.B.18.** (The Euler number) Combining the binomial theorem with the monotone sequence theorem, see 5.B.11, show that the sequence  $e_n$  given below is convergent:

$$e_n = \left( 1 + \frac{1}{n} \right)^n, \quad n = 1, 2, \dots$$

○

 Roughly speaking, Cauchy sequences are sequences whose all but a finite number of elements sit from each other less than a given distance  $\varepsilon > 0$ . A general result states that any convergent sequence is Cauchy, while over the reals  $\mathbb{R}$  or the complex numbers  $\mathbb{C}$  the converse is also true, see 5.2.3. Thus we say that  $\mathbb{R}$  or  $\mathbb{C}$  are complete fields, instead of  $\mathbb{Q}$ , for example. We can often use this result, to decide about the convergence/divergence of a given sequence of real or complex numbers. Let us describe such an example, and for more advanced tasks check the final section with the extra material, see 5.E.29, 5.E.30, for example.

<sup>5</sup>Good to know that the proof given here as a solution of 5.B.18, differs from those given in 5.4.1. Can you establish the different points between the two methods?

of those points. By this procedure, a Cauchy sequence is established. Cauchy sequences have limit points or are constant up to finitely many exceptions. Thus there is a subsequence with the desired limit. The fact that  $A$  is closed implies that the point obtained lies in  $A$ .

Now the other direction: if every infinite subset of  $A$  has a limit point in  $A$ , then all limit points are in  $A$ , and so  $A$  is closed. If  $A$  were not bounded, we would be able to find an increasing or decreasing sequence such that the differences of absolute values of adjacent numbers would be at least 1, for instance. However, such a sequence of points in  $A$  cannot have a limit point at all.

Finally, we have to deal with the general case  $A \subset \mathbb{C}$ . The arguments of the latter implication remain the same. Thus we have to show that any sequence  $z_n$  of complex numbers in  $A$  has got a limit point in  $A$ . Consider the sequences of real and imaginary parts,  $x_n$  and  $y_n$ . Since they both have to be in the bounded subsets  $A_{\mathbb{R}}$  and  $A_{i\mathbb{R}}$  of the real and imaginary projections of  $A$ , there is a subsequence  $z_{n_k} = (x_{n_k}, y_{n_k})$  such that  $x_{n_k} \rightarrow x$ ,  $y_{n_k} \rightarrow y$  with the limits sitting in the closures of  $A_{\mathbb{R}}$  and  $A_{i\mathbb{R}}$ , by virtue of the already proved real case. Obviously,  $z_{n_k} \rightarrow z = (x, y)$ , but the latter limit has to sit in  $A$  since  $A$  is closed.

(5) First, focus on the easier implication. That is, suppose that every open cover contains a finite subcover. It is required to prove that  $A$  is both closed and bounded.  $A \subset \mathbb{C}$  can be covered by a countable union of neighbourhoods  $\mathcal{O}_n(z)$ , with integers  $n$  and centers  $z$  with integral real and imaginary parts. Any choice of a finite subcover of them witnesses that  $A$  is bounded. If  $A \subset \mathbb{R}$ , then the same argument applies with intervals  $\mathcal{O}_n(x)$ ,  $n, x \in \mathbb{Z}$ .

Now suppose that  $a \in \mathbb{C} \setminus A$  is the limit point of a sequence  $a_i \in A$ . Further, assume that  $|a - a_n| < \frac{1}{n}$  (otherwise select a subsequence satisfying this property). The sets

$$J_n = \mathbb{C} \setminus \overline{\mathcal{O}_{1/n}(a)}$$

for all  $n \in \mathbb{N}$ ,  $n > 0$ , are open and they also cover our set  $A$ . Since it is possible to choose a finite cover of  $A$ , the point  $a$  is inside the complement  $\mathbb{C} \setminus A$ , including one of its neighbourhoods, and thus it is not a limit point. Therefore, all of  $A$ 's limit points must again lie in  $A$ . Hence  $A$  is closed.

If  $A \subset \mathbb{R}$ , the same argument applies with closed discs replaced by closed intervals.

Finally, we have to prove the other implication. So assume  $A \subset \mathbb{C}$  is complete and bounded, but there is an open covering  $U_\alpha$ ,  $\alpha \in I$ , of  $A$ , which does not contain any finite covering. Consider the sequence of positive real numbers  $\varepsilon_n = 1/n$  converging to 0 and sets

$$B_n = \{z = (\frac{k}{n}, \frac{m}{n}) \in A, k, m \in \mathbb{Z}\}$$

of complex numbers with real and imaginary parts in the “ $1/n$ -net of coordinates”. Clearly all sets  $B_n$  are finite. Further, for each  $k$ , consider the system  $\mathcal{A}_k$  of closed discs with centres in the points of  $B_k$  and diameters  $2\varepsilon_k$ . Clearly each such system  $\mathcal{A}_k$  covers the entire set  $A$ . Altogether, there

**5.B.19. Harmonic numbers.** Consider the sequence  $(H_n)_{n=1}^\infty$  with general term

$$H_n = \sum_{k=1}^n \frac{1}{k}, \quad n = 1, 2, \dots$$

Decide about the convergence of  $(H_n)_{n=1}^\infty$  in terms of Cauchy sequences.

**Solution.** The term  $H_n$  is the so-called *n*th harmonic number. The first harmonic numbers are given by

$$H_1 = 1, \quad H_2 = \frac{3}{2} = H_1 + \frac{1}{2}, \quad H_3 = \frac{11}{6} = H_2 + \frac{1}{3},$$

and by induction over  $n$  we can show that  $H_{n+1} = H_n + \frac{1}{n+1}$ , for all naturals  $n \geq 1$ . Harmonic numbers are the partial sums of the *harmonic series* that we will meet later in Section D. Now, for convenience set  $b = |H_{2n} - H_n|$ . Then

$$\begin{aligned} b &= \left| 1 + \frac{1}{2} + \dots + \frac{1}{n} + \frac{1}{n+1} + \dots + \frac{1}{2n} - H_n \right| \\ &= \frac{1}{n+1} + \frac{1}{n+2} + \dots + \frac{1}{2n} \\ &> \underbrace{\frac{1}{2n} + \frac{1}{2n} + \dots + \frac{1}{2n}}_{n\text{-terms}} = \frac{1}{2}. \end{aligned}$$

Thus taking  $\varepsilon = 1/2$  we see that there is no positive integer  $N$  satisfying  $|H_m - H_n| < 1/2$ , for all  $m, n \geq N$ . This means that  $(H_n)$  is not a Cauchy sequence. Thus  $(H_n)$  diverges (to  $+\infty$  since all of its terms are positive). We may verify the situation in Sage, by giving the cell

```
var("n");sum(1/n, n, 1, oo)
```

Sage after returning some “Runtime Errors”, types

```
ValueError: Sum is divergent.
```

□



Let us now present exercises having a bit more topological character, and are about limit, interior, boundary or isolated points of certain subsets of  $\mathbb{R}$  or  $\mathbb{C}$ . Here one should refresh the formal definitions given in 5.2.5, 5.2.7, and read carefully the proof of the Theorem 5.2.8. Mastering these notions may needs some effort, and thus we have included a series of exercises (see also the problems 5.E.36, 5.E.37, 5.E.38, and 5.E.39 in Section E).

**5.B.20. Limit, interior and isolated points.** Find all limit, isolated, and interior points of the set  $\mathbb{N}^* \subset \mathbb{R}$  of non-zero natural numbers (i.e., positive integers), and of the set  $\mathbb{Q} \subset \mathbb{R}$  of rational numbers.

**Solution.** Any  $n \in \mathbb{N}^*$  admits a neighbourhood in  $\mathbb{R}$  containing only the number  $n$ , i.e.,  $\mathcal{O}_1(n) \cap \mathbb{N}^* = (n-1, n+1) \cap \mathbb{N}^* = \{n\}$ . This means that every point  $n \in \mathbb{N}^*$  is isolated, and there are no interior points (since an isolated point cannot be an interior point). Moreover  $\mathbb{N}^*$  provides an example of an infinite set having no limits points. Indeed, if  $x_0$  is a

must be at least one closed disc  $C$  in the system  $\mathcal{A}_1$  which is not covered by a finite number the sets  $U_\alpha$ . Call it  $C_1$  and notice that  $\text{diam } C_1 = 2\varepsilon_1$ .

Next, consider the sets  $C_1 \cap C$ , with discs  $C \in \mathcal{A}_2$  which cover the entire set  $C_1$ . Again, at least one of them cannot be covered by a finite number of  $U_\alpha$ , we call it  $C_2$ . This way, we inductively construct a sequence of sets  $C_k$  satisfying  $C_{k+1} \subset C_k$ ,  $\text{diam } C_k \leq 2\varepsilon_k$ ,  $\varepsilon_k \rightarrow 0$ , and none of them can be covered by a finite number of the open sets  $U_\alpha$ .

Finally we choose one point  $z_k \in C_k$  in each of these sets. By construction, this must be a Cauchy-sequence. Consequently, this sequence of complex numbers has a limit  $z$ . Thus there is  $U_{\alpha_0}$  containing  $z$  and containing also some  $\delta$ -neighbourhood  $\mathcal{O}_\delta(z)$ . But now, if  $\text{diam } C_k \leq 2\varepsilon_k < \delta$ , then  $C_k \subset \mathcal{O}_\delta(z) \subset U_{\alpha_0}$ , which is a contradiction. The proof is complete when considering  $A \subset \mathbb{C}$ .

Dealing with real subset  $A \subset \mathbb{R}$ , again the same line of arguments applies, just the 2-dimensional nets  $B_k$  become 1-dimensional and the open discs are replaced by open intervals. □

**5.2.9. Limits of functions and sequences.** For the discussion of limits, it is advantageous to extend the set  $\mathbb{R}$  of real numbers by the two infinite values  $\pm\infty$  as we have done when defining intervals.

A neighbourhood of infinity is any interval  $(a, \infty)$ . Similarly, any interval  $(-\infty, a)$  is a neighbourhood of  $-\infty$ . Further, we will extend the concept of a limit point so that  $\infty$  is a limit point of a set  $A \subset \mathbb{R}$  if and only if every neighbourhood of  $\infty$  has a non-empty intersection with it, i.e. if the set  $A$  is unbounded from above. Similarly for  $-\infty$ . We talk about the infinite limit points, sometimes also called *improper limit points* of the set  $A$ .



“CALCULATIONS” WITH INFINITIES

We also introduce rules for calculation with the formally added values  $\pm\infty$  and arbitrary “finite” numbers  $a \in \mathbb{R}$ :

$$\begin{aligned} a + \infty &= \infty \\ a - \infty &= -\infty \\ a \cdot \infty &= \infty, \text{ if } a > 0 \\ a \cdot \infty &= -\infty, \text{ if } a < 0 \\ a \cdot (-\infty) &= -\infty, \text{ if } a > 0 \\ a \cdot (-\infty) &= \infty, \text{ if } a < 0 \\ \frac{a}{\pm\infty} &= 0, \text{ for all } a \neq 0. \end{aligned}$$

The following definition covers many cases of limit processes and needs to be thoroughly understood. Some particular cases are considered in great detail below.



limit point of  $\mathbb{N}^*$  then for some  $\varepsilon > 0$  we should have a natural (and in particular an infinite number of natural numbers) in distance smaller than  $\varepsilon$ , which is impossible.

Now for  $\mathbb{Q}$ , recall by 1.A.1 that this is a *dense* subset of  $\mathbb{R}$ . This means that each neighbourhood  $\mathcal{O}_\varepsilon(x)$  of any  $x \in \mathbb{R}$  contains rational numbers, that is for any  $x \in \mathbb{R}$  and every  $\varepsilon > 0$  there exists  $q \in \mathbb{Q}$  with  $q \in \mathcal{O}_\varepsilon(x) = (x - \varepsilon, x + \varepsilon)$ . Thus, for every  $x \in \mathbb{R}$ , there is a sequence of rational numbers  $a_n \neq x$ , converging to it. For instance, imagine the decimal representation of a real number  $x \notin \mathbb{Q}$ , and the corresponding sequence whose  $k$ th term will be the representation truncated to the first  $k$  decimal digits. If  $x$  was rational, we could clearly consider the sequence  $a_n = x + \frac{1}{n}$ . The set of all limit points of  $\mathbb{Q}$  is thus the whole real line  $\mathbb{R}$ . There are no isolated points in  $\mathbb{Q}$ , and in particular any rational number is also a limit point of the complement  $\mathbb{R} \setminus \mathbb{Q}$ . Finally, there are no interior points of  $\mathbb{Q}$  (this is because between any two rational numbers, there is an irrational number, and so it is impossible for a rational  $a \in \mathbb{Q}$  to be an interior point of  $\mathbb{Q}$ ).  $\square$

**5.B.21.** Find all limit, isolated, boundary and interior points of the sets  $X = \{x \in \mathbb{R} : 0 \leq x < 1\} \subset \mathbb{R}$  and  $Y = \{z \in \mathbb{C} : 0 \leq |z| < 1\} \subset \mathbb{C}$ .

**Solution.** Let  $a \in [0, 1)$  be an arbitrary number. Apparently, the sequences  $\{a + \frac{1}{n}\}_{n=1}^\infty$ ,  $\{1 - \frac{1}{n}\}_{n=1}^\infty$  converge to  $a$  and 1, respectively. So we have easily shown that the set of  $X$ 's limit points contains the interval  $[0, 1]$ . There are no other limit points: for any  $b \notin [0, 1]$  there is  $\delta > 0$  such that  $\mathcal{O}_\delta(b) \cap [0, 1] = \emptyset$  (for  $b < 0$  it suffices to take  $\delta = -b$ , and for  $b > 1$  we can choose  $\delta = b - 1$ ). Since every point of the interval  $[0, 1)$  is a limit point, there are no isolated points. For  $a \in (0, 1)$ , let  $\delta_a$  be the smaller one of the two positive numbers  $a$ ,  $1 - a$ . Considering  $\mathcal{O}_{\delta_a}(a) = (a - \delta_a, a + \delta_a) \subseteq (0, 1)$  for  $a \in (0, 1)$ , we see that every point of the interval  $(0, 1)$  is an interior point of  $X$ . For every  $\delta \in (0, 1)$ , we have that

$$\begin{aligned} \mathcal{O}_\delta(0) \cap [0, 1) &= (-\delta, \delta) \cap [0, 1) = [0, \delta), \\ \mathcal{O}_\delta(1) \cap [0, 1) &= (1 - \delta, 1 + \delta) \cap [0, 1) = (1 - \delta, 1), \end{aligned}$$

so every  $\delta$ -neighborhood of the point 0 contains some points of the interval  $[0, 1)$  and some points of the interval  $(-\delta, 0)$ , and every  $\delta$ -neighborhood of 1 has a non-empty intersection with the intervals  $[0, 1)$ ,  $[1, 1 + \delta)$ . Therefore, 0 and 1 are boundary points. Altogether, we have found that the interior points of  $X$  coincide with the interval  $(0, 1)$  while the two-element set  $\{0, 1\}$  consists of the boundary points of  $X$  (as we know that no point can be both interior and boundary and that a boundary point must be an isolated or a limit point). The case of  $Y$  is very similar and we leave most of the details to the reader. It is the open unit disk in the complex plane, all its points are interior (as with all open sets). The boundary points form the unit circle. Thus, the set of limit points is the closed unit disc and there are no isolated points there.  $\square$

**Definition.** Consider a subset  $A \subset \mathbb{R}$  and a real-valued function  $f : A \rightarrow \mathbb{R}$  or a complex-valued function  $f : A \rightarrow \mathbb{C}$ , defined on  $A$ . Further, consider a limit point  $x_0$  of the set  $A$  (i.e. a real number or  $\pm\infty$ ).

We say that  $f$  has *limit*  $a \in \mathbb{R}$  (or a complex limit  $a \in \mathbb{C}$ ) at  $x_0$  and write

$$\lim_{x \rightarrow x_0} f(x) = a$$

if and only if for every neighbourhood  $\mathcal{O}(a)$  of the point  $a$ , there is a neighbourhood  $\mathcal{O}(x_0)$  of  $x_0$  such that for all  $x \in A \cap (\mathcal{O}(x_0) \setminus \{x_0\})$ ,  $f(x) \in \mathcal{O}(a)$ .

In the case of a real-valued function,  $a = \pm\infty$  can also be the limit. Such a limit is called *infinite* or *improper*. In the other case, i.e.  $a \in \mathbb{R}$ , we say the limit is *finite* or *proper*.

It is important to notice that the value of  $f$  at  $x_0$  does not occur in the definition, and that the function  $f$  may even not be defined at this limit point (and in the case of an improper limit point, it cannot be defined, of course!).

We shall not deal with improper limits of complex functions now.

**5.2.10. The most important cases of domains.** Our definition of a limit covers several very dissimilar situations:

(1) **Limits of sequences.** If  $A = \mathbb{N}$ , i.e. the function  $f$  is defined for the natural numbers only, we talk about limits of sequences of real or complex numbers. In this case, the only limit point of the domain is  $\infty$ , and we mostly write the values (terms) of the sequence as  $f(n) = a_n$  and the limit in the form

$$\lim_{n \rightarrow \infty} a_n = a.$$

According to the definition, this means that for any neighbourhood  $\mathcal{O}(a)$  of the limit value  $a$ , there is an index  $N \in \mathbb{N}$  such that  $a_n \in \mathcal{O}(a)$  for all  $n \geq N$ . Actually, we have only reformulated the definition of convergence of a sequence (see 5.2.3). We have only added the possibility of infinite limits. As before, we also say that the *sequence*  $a_n$  *converges* to  $a$ .

We can easily see from our definition for complex numbers that a sequence of complex values has limit  $a$  if and only if the real parts of  $a_i$  converge to  $\operatorname{Re} a$  and the imaginary parts converge to  $\operatorname{Im} a$ .

(2) **Limits of functions at interior points of intervals.** If  $f$  is defined on the interval  $A = (a, b)$  and  $x_0$  is an interior point of this interval, we talk about the limit of a function at an interior point of its domain. Usually, we write

$$\lim_{x \rightarrow x_0} f(x) = a.$$

Let us examine why it is important to require  $f(x) \in \mathcal{O}(a)$  only for the points  $x \neq x_0$  in this case as well. As an example, let us consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$

$$f(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ 1 & \text{if } x = 0. \end{cases}$$

*If all limit points of a subset  $A$  of  $\mathbb{R}$  or of  $\mathbb{C}$  belong to  $A$ , then the subset is said to be closed. This means that we cannot run away from  $A$  by a sequence  $(x_n) \in A$  to a limit not belonging to  $A$ . If the complement of  $A$  in the real line or in the complex plane is closed, then clearly each  $x \in A$  has a neighbourhood  $\mathcal{O}_\varepsilon(x) \subset A$ , for some  $\varepsilon > 0$ . In this case  $A$  is called open, see 5.2.5, 5.2.7 and 5.2.8 for further details. Observe that the empty set  $\emptyset$  and  $\mathbb{R}$  (or  $\mathbb{C}$ ) itself are both open and closed (for instance,  $\mathbb{C}$  has no boundary points). If a subset  $A$  in  $\mathbb{R}$  or  $\mathbb{C}$  is closed and bounded (the latter means that  $A \subset \mathcal{O}_r(x)$  for some  $x$  and big enough  $r > 0$ ), then each sequence  $(x_n) \in A$  has a limit point in  $A$ , i.e., there is a convergent subsequence in  $A$ . Such sets are called compact sets.*

**5.B.22. Closed and open subsets.** Verify that an open interval  $I = (a, b) = \{x \in \mathbb{R} : a < x < b\}$  of the real line is an open set, and moreover that a closed interval  $J = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$  is a closed set. Is the interval  $(0, 1]$  a closed set?

**Solution.** Given some  $x \in I$  we should show that  $x$  is an interior point of  $I$ . Set  $\varepsilon = \frac{1}{2} \min\{x - a, b - x\} > 0$  and note that  $\varepsilon$  depends on  $x$ . Then we have  $a \leq x - \varepsilon < x + \varepsilon \leq b$ , and so  $(x - \varepsilon, x + \varepsilon) \subset (a, b)$ . Therefore, we have constructed an open neighbourhood  $\mathcal{O}_\varepsilon(x) \subset I$  around  $x$ , contained in  $I$ , and since this works for any  $x \in I$ , the set  $I$  is open. Next, the complement of  $J$  is  $J^c = (-\infty, a) \cup (b, \infty)$  which is open as a union of open intervals. Thus  $J$  is closed. Finally, the interval  $(0, 1]$  cannot be a closed set; For instance, the sequence  $(x_n) = (1/n) \in (0, 1]$  satisfies  $(1/n) \rightarrow 0$  as  $n \rightarrow \infty$ , but  $0 \notin (0, 1]$ .  $\square$

**5.B.23.** Decide which of the following sets are open or closed:

- (a)  $A = \mathbb{Z} \subset \mathbb{R}$ ;
- (b)  $B = [2, +\infty) \subset \mathbb{R}$ ;
- (c)  $C = \{z \in \mathbb{C} : |z - 1| + |z + 1| < 5\} \subset \mathbb{C}$ .

**Solution.** (a) It is not hard to see that the complement of  $\mathbb{Z}$  in  $\mathbb{R}$  is given by the union  $\cup_{k \in \mathbb{Z}} (k, k + 1)$ . Hence the set  $\mathbb{R} \setminus \mathbb{Z}$  is open, as the union of open sets. It follows that  $\mathbb{Z}$  is closed in  $\mathbb{R}$ .

(b) The complement of  $B$  in  $\mathbb{R}$  is the set  $(-\infty, 2)$ , which is open. Thus  $B$  itself is a closed subset of  $\mathbb{R}$ . In a similar way one can show that the following sets are closed:  $[a, +\infty)$ , and  $(-\infty, b]$ , with  $a, b \in \mathbb{R}$ .

(c) Let us treat this case based on the notion of continuity, that we will essentially analyze a few below. Consider the function  $f : \mathbb{C} \rightarrow \mathbb{R}$  given by  $f(z) = |z - 1| + |z + 1|$ . As a challenge for you, prove that this function is continuous. Moreover, show that  $C = f^{-1}((-\infty, 5))$ . Then, the claim follows by 5.2.18; The set  $(-\infty, 5)$  is open and  $f$  is continuous, hence  $C$  is open, as well.  $\square$

**5.B.24.** Is the range of  $f(x) = x^2$  a closed subset of  $\mathbb{R}$ ?  $\circ$

Apparently, the limit at zero is well-defined, and in accordance with our expectations,  $\lim_{x \rightarrow 0} f(x) = 0$  even though the value  $f(0) = 1$  does not belong into small neighbourhoods of the limit value 0.

An equivalent definition using  $\varepsilon$ -neighbourhoods of the limits  $a$  and  $\delta$ -neighbourhoods of the limit points  $x_0$  is the following:  $\lim_{x \rightarrow x_0} f(x) = a$  if for each  $\varepsilon > 0$  there is a  $\delta > 0$  such, that for all  $x \neq x_0$  satisfying  $|x - x_0| < \delta$ ,  $|f(x) - a| < \varepsilon$ .

(3) **One-sided limits.** If  $A = [a, b]$  is a bounded interval and  $x_0 = a$  or  $x_0 = b$ , we talk about a one-sided limit of the function  $f$  at the point  $x_0$ , from the right and from the left respectively.

If the point  $x_0$  is an interior point of the domain of  $f$ , we can, in order to determine the limit, consider the domain restricted to  $[x_0, b]$  or  $[a, x_0]$ . The resulting limits are also called a *right-sided limit* and *left-sided limit*, respectively, of the function  $f$  at the point  $x_0$ . We denote them by  $\lim_{x \rightarrow x_0^+} f(x)$  and  $\lim_{x \rightarrow x_0^-} f(x)$ , respectively. As an example, we can consider the one-sided limits at  $x_0 = 0$  for Heaviside's function  $h$  from the introduction to this part. Apparently,

$$\lim_{x \rightarrow 0^+} h(x) = 1, \quad \lim_{x \rightarrow 0^-} h(x) = 0.$$

However, the limit  $\lim_{x \rightarrow 0} f(x)$  does not exist.

It follows from the definitions that *the limit at an interior point of the domain of an arbitrary function  $f$  exists if and only if both one-sided limits exist and are equal.*

**5.2.11. Further examples of limits.** (1) The limit of a complex function  $f : A \rightarrow \mathbb{C}$  in a limit point  $x_0$  of its domain exists if and only if the limits of both the real part and the imaginary part exist. In this case, we have

$$\lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} (\operatorname{Re} f(x)) + i \lim_{x \rightarrow x_0} (\operatorname{Im} f(x)).$$

The proof is straightforward and makes direct use of the definitions of distances and neighbourhoods of the points in the complex plane. Indeed, the membership into a  $\delta$ -neighbourhood of a complex value  $z$  is guaranteed by the real  $(1/\sqrt{2})\delta$ -neighbourhoods of the real and the imaginary parts of  $z$ . Hence the proposition follows immediately.

(2) Let  $f$  be a real or complex polynomial. Then for every point  $x \in \mathbb{R}$ ,

$$\lim_{x \rightarrow x_0} f(x) = f(x_0).$$

Really, if  $f(x) = a_n x^n + \dots + a_0$ , then the identity  $(x_0 + \delta)^k = x_0^k + k\delta x_0^{k-1} + \dots + \delta^k$ , substituted for  $k = 0, \dots, n$ , gives that choosing a sufficiently small  $\delta$  makes the values arbitrarily close to  $f(x_0)$ .

(3) Now consider the following function defined on the whole real line

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$

It is apparent from the definition that this function cannot have (even one-sided) limits at any point of its domain.

**5.B.25.** Decide which of the following sets is compact:

- (a)  $I = (0, 1)$ ;
- (b)  $J = [0, 1]$ ;
- (c)  $A = \mathbb{N}^*$ ;
- (d)  $B = \{1/n : n \in \mathbb{N}^*\}$ ;
- (e)  $C = B \cup \{0\}$ .

**Solution.** Recall by (4) in Theorem 5.2.8 (Bolzano-Weierstrass theorem), that a subset  $A \subset \mathbb{R}$  is compact if and only if every finite sequence contained in  $X$  has a subsequence converging to a point in  $X$ . We can use the Bolzano-Weierstrass theorem as criterium to decide which of the given sets is compact or not.

- (a) The open set  $I = (0, 1)$  is not compact: Consider for example the sequence  $(x_n) = (\frac{1}{n})$ , as above. We know that in  $I$  this sequence converges to 0, but  $0 \notin I$ . Thus, for  $(x_n)$  there is no convergent subsequence whose limit belongs to  $I$ .
- (b) The closed set  $J = [0, 1]$  is compact, as every bounded closed interval  $[a, b] \subset \mathbb{R}$ , see 5.E.39.
- (c) The set of non-zero naturals  $\mathbb{N}^*$  is not compact. For instance, the sequence  $(y_n = n)$  has no convergent subsequence. This is because every subsequence diverges to infinity.
- (d) The set  $B$  is not closed, so cannot be compact either.
- (e) Obviously, the set  $C$  is closed and bounded, and hence compact.  $\square$

Given a sequence  $(a_n)$  in 5.4.6 we consider the so called “upper limit” and “lower limit”, also called *limes superior* and *inferior* of  $(a_n)$ , and denoted by  $\limsup_{n \rightarrow \infty} a_n$  and  $\liminf_{n \rightarrow \infty} a_n$ . These limits always exists and have many important applications (see for example in Section D), and see also 5.E.32 and 5.E.34 for more details.

**5.B.26.** Compute the limes superior and inferior of the sequence

$$a_n = \frac{n^2 + 4n - 5}{n^2 + 9} \sin^2\left(\frac{n\pi}{4}\right), \quad n \in \mathbb{N}^*.$$

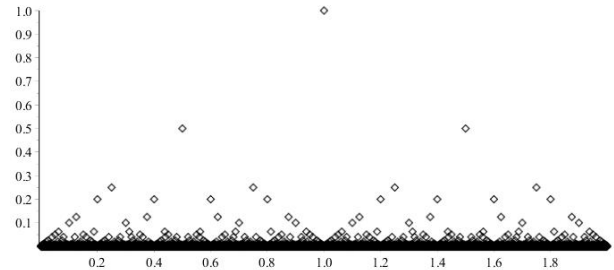
**Solution.** Recall that the limes superior/inferior are the largest/smallest limit point of the sequence. For the given task to compute them we may split the sequence  $(a_n)$  into several subsequences according the the value of  $(\sin \frac{n\pi}{4})^2$ , e.g., 0,  $\frac{1}{2}$ , and 1. All of them are converging to different limits and the result is  $\limsup_{n \rightarrow \infty} a_n = 1$  and  $\liminf_{n \rightarrow \infty} a_n = 0$ , respectively.  $\square$

Additional exercises on limes superior/inferior and most of the directions analyzed above are presented in Section E. Later, in Chapter 7, we will revise most of the topological notions analyzed in this chapter (using  $\mathbb{R}$  or  $\mathbb{C}$ ), in terms of the more general notion of “metric spaces”. But to do so, we first need to learn for limits of functions and continuous functions.



(4) The following function is even trickier than the previous one. Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be the function defined as follows:<sup>8</sup>

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{1}{q} & \text{if } x = \frac{p}{q} \in \mathbb{Q}, p, q \in \mathbb{Z}, q > 0 \text{ are co-prime,} \\ 0 & \text{if } x \notin \mathbb{Q}. \end{cases}$$



Choose any point  $x_0$ , no matter whether rational or irrational. Our goal is to show that  $\lim_{x \rightarrow x_0} f(x) = 0$ . Thus fix any  $\varepsilon > 0$  and look at the possible values of  $f(x)$  close to  $x_0$ . Notice that  $f(x) \geq \varepsilon$ , i.e.  $\frac{1}{q} \geq \varepsilon$  can be true for only finite number of  $q \in \mathbb{N}$ . This behaviour is illustrated on the diagram. In particular, there can be only a finite number of points  $x$  in the interval  $(x_0 - 1, x_0 + 1)$  for which  $f(x) \geq \varepsilon$ . Label them  $x_1, \dots, x_n$ . Finally, choose  $\delta$  smaller than the minimum of the distances of any two different points  $x_i$ . Then  $f(x) < \varepsilon$  for all  $x \in \mathcal{O}_\delta(x_0) \setminus \{x_0\}$ . This finishes the proof.



Notice that this limit equals the function value only at the irrational points.

**5.2.12. The squeeze theorem.** The following result is elementary, but extremely useful. We meet it when computing limits of all types discussed above, i.e. limits of sequences, limits of functions at interior points, one-sided limits, and so on.



**Theorem.** Let  $f, g, h$  be three real-valued functions with the same domain  $A$  and such that there is a neighbourhood of a limit point  $x_0 \in \mathbb{R}$  of the domain where

$$f(x) \leq g(x) \leq h(x),$$

for all  $x \neq x_0$ . Suppose there are limits

$$\lim_{x \rightarrow x_0} f(x) = f_0, \quad \lim_{x \rightarrow x_0} h(x) = h_0$$

and  $f_0 = h_0$ . Then the limit

$$\lim_{x \rightarrow x_0} g(x) = g_0$$

exists, and it satisfies  $g_0 = f_0 = h_0$ .

<sup>8</sup>This function is called the *Thomae function* after the German mathematician J. Thomae, 1840–1921. You may find it under many other names too: e.g. *Riemann function*, *pop-corn function*, *raindrop function* etc. It illustrates how badly dense the “discontinuity” points of a function can be even though it has limits everywhere.

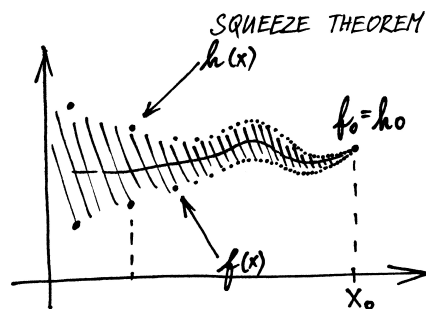
Limits of functions enable us to investigate the behaviour of functions (and hence their graph), in the neighbourhood of some given point. They form the basis of the notion of differentiation that we will discuss in the forthcoming section, and establishes one of the most beautiful themes in mathematical analysis and in many other areas (e.g., numerical analysis).

Most of the tasks presented below are based on theoretical results from the sections 5.2.9, 5.2.12, 5.2.13, 5.2.16, 5.2.18, and 5.2.20. All these results will gradually become more evident, as we explore them by examples. We recall however that limits are well defined in the limit points of the given domain of a function, and the function is called continuous at  $x$ , if its limit equals the function value there (be aware however, that in general a function does not need to be defined at the point where we seek for the limit). Recall also that from the topological point of view (see 5.2.18), a function  $f : A \subset \mathbb{R} \rightarrow \mathbb{R}$  is continuous if and only if the pre-image  $f^{-1}(U)$  is an open subset of  $A$ , for every open set  $U \subset \mathbb{R}$ . To verify that  $f^{-1}(U)$  is open it suffices to show that any point  $x \in f^{-1}(U)$  lies on an open interval of  $f^{-1}(U)$ .

**5.B.27. Limit of functions.** Find the following limits or explain why they do not exist:

- (a)  $\lim_{x \rightarrow c} x^2$ ,
- (b)  $\lim_{x \rightarrow c} (x^6 - x^3 - \sqrt{6})$ ,
- (c)  $\lim_{x \rightarrow 2} \frac{x^2 + x - 6}{x^2 - 3x + 2}$ ,
- (d)  $\lim_{x \rightarrow 0} \sin\left(\frac{1}{x}\right)$ ,
- (e)  $\lim_{x \rightarrow 2^+} \frac{x^2 + 2x - 8}{|x - 2|}$ ,
- (f)  $\lim_{x \rightarrow 2^-} \frac{x^2 + 2x - 8}{|x - 2|}$ ,
- (g)  $\lim_{x \rightarrow 3} \frac{x - 3}{\sqrt{x + 1} - (x - 1)}$ ,
- (k)  $\lim_{x \rightarrow 0} x \cos\left(\frac{1}{x}\right)$ .

**Solution.** (a) We can treat this limit by basic principles, as for example the Cauchy definition for limits of functions. Let us however rely on topology, as it was mentioned a few above, to show that  $f(x) = x^2$  is continuous and hence  $\lim_{x \rightarrow c} x^2 = c^2$ , where  $c$  is any real number. Let  $I = (a, b) \subset \mathbb{R}$  be an open interval with  $a < b$  and  $a, b \in \mathbb{R}$ . By definition,  $f^{-1}(I) = \{x \in \mathbb{R} : x^2 \in (a, b)\}$ , hence obviously for  $b \leq 0$  it should be  $f^{-1}(I) = \emptyset$  (since  $x^2 \geq 0$  for any  $x$ ), which is open. For  $a < 0 < b$ , one gets  $f^{-1}(I) = (-\sqrt{b}, \sqrt{b}) \subset \mathbb{R}$ , since any  $x \in (-\sqrt{b}, \sqrt{b})$  satisfies  $f(x) \in (a, b)$ . Thus again  $f^{-1}(I)$  is open as an open interval of  $\mathbb{R}$ . Finally, for  $0 \leq a < b$  the preimage  $f^{-1}(I)$  is the union  $(-\sqrt{b}, -\sqrt{a}) \cup (\sqrt{a}, \sqrt{b})$  which is again open as the union of two open intervals of  $\mathbb{R}$ . (b) For the second limit one simply applies sum and difference rules (see also below), and get benefit from the fact that the  $n$ th power  $x \mapsto x^n$  ( $n = 0, 1, 2, \dots$ ) is a continuous function. Or you can directly rely on the fact that polynomials are continuous everywhere in  $\mathbb{R}$ , see 5.2.11. Hence  $\lim_{x \rightarrow c} (x^6 - x^3 + x - \sqrt{6}) = c^6 - c^3 - \sqrt{6}$ . (c) The substitution  $x = 2$  leads to both zero numerator and zero denominator. Despite of that, the problem can be solved



**PROOF.** From the assumptions of the theorem, it follows that for any  $\varepsilon > 0$ , there is a neighbourhood  $\mathcal{O}(x_0)$  of the point  $x_0 \in A \subset \mathbb{R}$  in which both  $f(x)$  and  $h(x)$  lie in the interval  $(f_0 - \varepsilon, f_0 + \varepsilon)$ , for all  $x \neq x_0$ . From the condition  $f(x) \leq g(x) \leq h(x)$ , it follows that  $g(x) \in (f_0 - \varepsilon, f_0 + \varepsilon)$ , and so  $\lim_{x \rightarrow x_0} g(x) = f_0$ .

The above reasoning is easily modified for infinite limit values or for limits at infinite points  $x_0$ . In the first case, choose a large  $N$  instead of  $\varepsilon$ . The condition on the values reads: both  $f(x)$  and  $h(x)$  have values larger than  $N$  on the neighbourhood  $\mathcal{O}(x_0) \setminus \{x_0\}$ , and thus the same will be true for  $g(x)$ . In the second case, the neighbourhood  $\mathcal{O}$  will be an interval  $(M, \infty)$ . The other infinite limit point  $-\infty$  is dealt with similarly.  $\square$

The next theorem reveals the elementary properties of limits, again for all types together. Think about the individual cases, including the limits taken at  $x_0 = \pm\infty$ !

**5.2.13. Theorem.** Let  $A \subset \mathbb{R}$  be the domain of real or complex functions  $f$  and  $g$ , let  $x_0$  be a limit point of  $A$  and let the limits

$$\lim_{x \rightarrow x_0} f(x) = a \in \mathbb{K}, \quad \lim_{x \rightarrow x_0} g(x) = b \in \mathbb{K}$$

exist. Then:

- (1) the limit  $a$  is unique,
- (2) the limit of the sum  $f + g$  exists and satisfies

$$\lim_{x \rightarrow x_0} (f(x) + g(x)) = a + b,$$

- (3) the limit of the product  $f \cdot g$  exists and satisfies

$$\lim_{x \rightarrow x_0} (f(x) \cdot g(x)) = a \cdot b.$$

In particular, if  $f(x) = a$  is a constant function then  $\lim_{x \rightarrow x_0} a \cdot g(x) = a \cdot b$ ,

- (4) if  $b \neq 0$ , the limit of the quotient  $f/g$  exists and satisfies

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \frac{a}{b}.$$

**PROOF.** (1) Suppose  $a$  and  $a'$  are two values of the limit  $\lim_{x \rightarrow x_0} f(x)$ . If  $a \neq a'$ , then there are disjoint neighbourhoods  $\mathcal{O}(a)$  and  $\mathcal{O}(a')$ . However, for sufficiently small neighbourhoods of  $x_0$ , the values of  $f$  should lie in both neighbourhoods. This is a contradiction. Thus  $a = a'$ .

(2) Choose a neighbourhood of  $a+b$ , for instance  $\mathcal{O}_{2\varepsilon}(a+b)$ . For a sufficiently small neighbourhood of  $x_0$  and  $x \neq x_0$ ,

very easily by factorization:

$$\lim_{x \rightarrow 2} \frac{x^2 + x - 6}{x^2 - 3x + 2} = \lim_{x \rightarrow 2} \frac{(x-2)(x+3)}{(x-2)(x-1)} = \frac{2+3}{2-1} = 5.$$

(d) We will prove that this limit does not exist, using  $\varepsilon$ -neighbourhoods, see 5.2.9 and 5.2.10. Set  $f(x) = \sin(1/x)$  and assume, in contrast, that  $\lim_{x \rightarrow 0} f(x) = \ell$  for some real number  $\ell$ . This means that for  $\varepsilon = \frac{1}{2}$  there exists  $\delta > 0$  such that for any  $x$  with  $0 < |x - 0| < \delta$  we have  $|f(x) - \ell| < \frac{1}{2}$ . Then there exist some  $n \in \mathbb{N}$  and reals  $x_1 = \frac{1}{2n\pi}$  and  $x_2 = \frac{1}{2n\pi + \frac{\pi}{2}}$ , such that  $0 < |x_1| < \delta$  and  $0 < |x_2| < \delta$ . Hence,  $|f(x_1) - \ell| < \frac{1}{2}$ ,  $|f(x_2) - \ell| < \frac{1}{2}$ , and using the triangle inequality we see that

$$\begin{aligned} \Delta &= |(f(x_1) - \ell) + (\ell - f(x_2))| \\ &\leq |f(x_1) - \ell| + |\ell - f(x_2)| \\ &= |f(x_1) - \ell| + |f(x_2) - \ell| < \frac{1}{2} + \frac{1}{2} = 1, \end{aligned}$$

where for simplicity we set  $\Delta := |f(x_1) - f(x_2)|$ . Since  $f(x_1) = \sin(2n\pi) = 0$  and  $f(x_2) = \sin(2n\pi + \frac{\pi}{2}) = \sin(\frac{\pi}{2}) = 1$ , the last inequality gives  $|f(x_1) - f(x_2)| = |0 - 1| = 1 < 1$ , a contradiction.

(e) By the definition of the absolute value we have

$$\begin{aligned} \lim_{x \rightarrow 2^+} \frac{x^2 + 2x - 8}{|x - 2|} &= \lim_{x \rightarrow 2^+} \frac{(x-2)(x+4)}{x-2} \\ &= \lim_{x \rightarrow 2^+} (x+4) = 6. \end{aligned}$$

(f) Similarly

$$\lim_{x \rightarrow 2^-} \frac{x^2 + 2x - 8}{|x - 2|} = \lim_{x \rightarrow 2^-} \frac{(x-2)(x+4)}{-(x-2)} = -6.$$

Thus, a comparison with the result from case (e) implies that the function  $f(x) = \frac{x^2 + 2x - 8}{|x - 2|}$  cannot be continuous at  $x_0 = 2$  (since the one sided limits are not equal).

(g) Set for simplicity  $f(x) = \frac{x-3}{\sqrt{x+1} - (x-1)}$ . When  $x \rightarrow 3$  we see that both the numerator and denominator tend to 0, which gives the indeterminate form  $\frac{0}{0}$  (see also below). One can treat such cases by multiplying both the numerator and denominator by the “conjugate expression” of the denominator. For our case this means  $\sqrt{x+1} + (x-1)$  with

$$(\sqrt{x+1} - (x-1))(\sqrt{x+1} + (x-1)) = -x(x-3).$$

Thus

$$\begin{aligned} \lim_{x \rightarrow 3} f(x) &= \lim_{x \rightarrow 3} \frac{(x-3)(\sqrt{x+1} + (x-1))}{-x(x-3)} \\ &= -\lim_{x \rightarrow 3} \frac{\sqrt{x+1} + (x-1)}{x} = -\frac{4}{3}. \end{aligned}$$

(k) Recall that  $|\cos(x)| \leq 1$  for any  $x \in \mathbb{R}$  and hence  $|x \cos(\frac{1}{x})| \leq |x|$ . By the squeeze theorem (5.2.12) this gives  $\lim_{x \rightarrow 0} x \cos(\frac{1}{x}) = 0$ .  $\square$

both  $f(x)$  and  $g(x)$  will lie in  $\varepsilon$ -neighbourhoods of the points  $a$  and  $b$ . Hence their sum will lie in the  $2\varepsilon$ -neighbourhood of  $a + b$ . The proposition is proved.

(3) We have to be a bit more careful here. Let us look, what we can do with an assumption  $|f(x) - a| < \varepsilon$ ,  $|g(x) - b| < \varepsilon$  for  $x \in \mathcal{O}(x_0)$ , i.e. choosing a small enough neighbourhood of the limit point  $x_0$ . Estimate:

$$\begin{aligned} |f(x)g(x) - ab| &= |f(x)(g(x) - b) + b(f(x) - a)| \\ &\leq |f(x)|\varepsilon + |b|\varepsilon \leq |f(x) - a|\varepsilon + |a|\varepsilon + |b|\varepsilon \\ &\leq \varepsilon^2 + \varepsilon(|a| + |b|) \end{aligned}$$

Now we easily conclude. Choosing any  $\tilde{\varepsilon} > 0$ , there is unique  $\varepsilon > 0$  with  $\tilde{\varepsilon} = \varepsilon^2 + \varepsilon(|a| + |b|)$ . Thus using this  $\varepsilon$  for the choice of  $\mathcal{O}(x_0)$  above, we arrive at the required condition from the definition of the limit.

Clearly, the limit of the constant function  $f(x) = a$  is  $a$  at all limit points of its domain.

(4) In view of the previous results, it suffices to prove  $\lim_{x \rightarrow x_0} \frac{1}{g(x)} = \frac{1}{b}$  for  $|b| > 0$ . We need to be careful when considering complex valued functions. We need to estimate

$$\left| \frac{1}{g(x)} - \frac{1}{b} \right| = \left| \frac{b - g(x)}{g(x) \cdot b} \right|.$$

Since  $|b| > 0$ , we may restrict ourselves to a neighbourhood  $U$  of  $x_0$  such that  $|g(x)| > \frac{|b|}{2}$ . Then  $|g(x)| \cdot |b| > \frac{|b|^2}{2}$ . Thus, if  $|g(x) - b| < \varepsilon$ , then

$$\left| \frac{1}{g(x)} - \frac{1}{b} \right| < \frac{2|b - g(x)|}{|b|^2} < \frac{2\varepsilon}{|b|^2}.$$

This verifies the claim as in the previous case.  $\square$

**5.2.14. Remarks on infinite values of limits.** The statement of the theorem can be extended to some infinite values of the limits of real-valued functions: For sums, either at least one of the two limits must be finite or both limits share the same sign. Then the limit of the sum is the sum of the limits, with the conventions from 5.2.9. However, “ $\infty - \infty$ ” is excluded.

For products, if one of the limits is infinite, then the other limit must be non-zero. Then the limit of the product is the product of the limits. The case “ $0 \cdot (\pm\infty)$ ” is excluded.

For a quotient, it may be that  $a \in \mathbb{R}$  and  $b = \pm\infty$ , then the resulting limit will be zero; or  $a = \pm\infty$  and  $b \in \mathbb{R}$ , then it will be  $\pm\infty$  according to the signs of the numerator and the denominator. The case “ $\frac{\infty}{\infty}$ ” is excluded.

The theorem also covers, as a special case, the corresponding statements about the convergence of sequences as well as about one-sided limits of functions defined on an interval.

The following provides a “convergence test” useful in many situations. It relates to limits of sequences and functions in general.

**5.2.15. Proposition.** Consider a real or complex valued function  $f$  defined on a set  $A \subset \mathbb{R}$  and a limit point  $x_0$  of the set  $A$ .  $f$  has a limit  $y$  at  $x_0$  if and only if for every sequence of



**5.B.28.** Compute the limit

$$\lim_{x \rightarrow 1} \frac{3x^3 + 2x^2 - 2x - 3}{5x^3 + 2x^2 - 2x - 5},$$

if it is known that  $\rho = 1$  is a root of both the numerator and denominator. Next use Sage to confirm your computations.

**Solution.** First notice that the function  $f(x) = \frac{3x^3+2x^2-2x-3}{5x^3+2x^2-2x-5}$  is not defined at  $x = 1$ , and moreover this limit has again the indeterminate form  $\frac{0}{0}$ . However, the statement already motivates a method for solving the task: Use the Horner's scheme and factorize the numerator and the denominator, i.e.,

$$\begin{aligned} 3x^3 + 2x^2 - 2x - 3 &= (x - 1)(3x^2 + 5x + 3), \\ 5x^3 + 2x^2 - 2x - 5 &= (x - 1)(5x^2 + 7x + 5). \end{aligned}$$

Then it is easy to see that the limit at hand equals to  $11/17$ .

In fact, you can easily do all the computations that we omitted via Sage, where you can still use the command `limit` to compute limits of functions, as in the case of sequences. Here it is however necessary to specify the limit point  $x_0$ , and the exact syntax for computing the limit  $\lim_{x \rightarrow x_0} f(x)$  is `limit(f(x), x = x_0)`. In addition, you can program Sage to compute the left and right side limits, by adding the options `dir = "left"` and `dir = "right"`, respectively. For our task the cell

```
g(x)=3*x^3+2*x^2-2*x-3; h(x)=5*x^3+2*x^2-2*x-5
show(factor(g(x))); show(factor(h(x)))
limit(g(x)/h(x), x=1)
```

verifies the factorizations given above and the explicit value of the limit, as you can easily check in your computer.

Finally, it is worth mentioning that one could compute the limit at hand by other methods. For instance, one can apply Proposition 5.2.15. Hence, let  $(x_n)$  be a sequence of real numbers with  $x_n \rightarrow 1$  as  $n \rightarrow \infty$ , and  $x_n \neq 1$  for all  $n = 1, 2, \dots$ . Then, for the induced sequence of values  $f(x_n)$  we see that

$$\begin{aligned} f(x_n) &= \frac{3x_n^3 + 2x_n^2 - 2x_n - 3}{5x_n^3 + 2x_n^2 - 2x_n - 5} \\ &= \frac{(x_n - 1)(3x_n^2 + 5x_n + 3)}{(x_n - 1)(5x_n^2 + 7x_n + 5)} = \frac{3x_n^2 + 5x_n + 3}{5x_n^2 + 7x_n + 5}. \end{aligned}$$

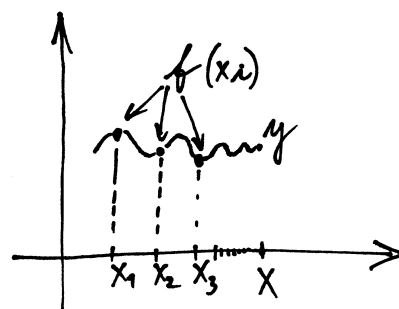
Taking the limit  $\lim_{n \rightarrow \infty} f(x_n)$  we get  $11/17$ , and according to Proposition 5.2.15 this implies that  $\lim_{x \rightarrow 1} f(x) = 11/17$ .  $\square$

**5.B.29.** Explain why the relation  $\lim_{x \rightarrow \pi/3} \sin(x) = \sin(\pi/3)$  is true. Next use Sage to confirm that  $\sin(x)$  is continuous at the chosen point by combining the `limit` command with the `bool` syntax.  $\circ$

*The Cauchy definition of limits of functions in terms of  $\varepsilon$  and  $\delta$  can be used to derive a number of elementary results. Let us describe such an example.*

points  $x_n \in A$  converging to  $x_0$ ,  $x_n \neq x_0$ , the sequence of the values  $f(x_n)$  has limit  $y$ .

### CONVERGENCE TEST



**PROOF.** Suppose first that the limit of  $f$  at  $x_0$  is  $y$ . Then for any neighbourhood  $U$  of the point  $y$ , there is a neighbourhood  $V$  of  $x_0$  such that for all  $x \in V \cap A$ ,  $x \neq x_0$ ,  $f(x) \in U$ . For every sequence  $x_n \rightarrow x_0$  of points different from  $x_0$ , the terms  $x_n$  lie in  $V$  for all  $n$  greater than a suitable  $N$ . Therefore, the sequence  $f(x_n)$  converges to  $y$ .

Now suppose that the function  $f$  does not converge to  $y$  at  $x \rightarrow x_0$ . Then for some neighbourhood  $U$  of  $y$ , there is a sequence of points  $x_m \neq x_0$  in  $A$  which are closer to  $x_0$  than  $1/m$ , with  $f(x_m)$  not belonging to  $U$ . In this way, there is constructed a sequence of points lying in  $A$  different from  $x_0$ , with  $\lim_{m \rightarrow \infty} x_m = x_0$ , for which the values  $f(x_n)$  do not converge to  $y$ . The proof is finished.  $\square$

**5.2.16. Continuity.** Continuity was discussed intuitively when polynomials were discussed. Now all the tools for a proper formulation of continuity are prepared. This is the basic class of functions in the sequel.



### CONTINUITY OF FUNCTIONS

**Definition.** Let  $f$  be a real or complex valued function defined on an interval  $A \subset \mathbb{R}$ .  $f$  is *continuous at a point*  $x_0 \in A$  if and only if

$$\lim_{x \rightarrow x_0} f(x) = f(x_0).$$

The function  $f$  is continuous on an interval  $A$  if and only if it is continuous at every point  $x_0 \in A$ .

The diagram explains the meaning of continuity. Firstly, the limit has to exist. Thus, after choosing a neighbourhood  $U$  of the limit value  $f(x_0)$  (the  $\varepsilon$ -neighbourhood  $\mathcal{O}_\varepsilon(f(x_0))$  is shown), there is a neighbourhood of  $x_0$  (the  $\delta$ -neighbourhood is shown), for which all images lie in  $U$ . In words, if we decide how close we want to be to  $f(x_0)$ , we always may choose a sufficiently small neighbourhood of  $x_0$  where this is guaranteed.

**5.B.30.** Let  $f : A \rightarrow \mathbb{R}$  be a function, where  $A$  is a subset of  $\mathbb{R}$  and let  $a$  be a limit point of  $A$ . Suppose that the limit  $\lim_{x \rightarrow a} f(x)$  exists and is a finite number. Show that  $f$  is bounded on the set  $(a - h, a + h) \cap A$  for some  $h > 0$ .

**Solution.** By assumption, the limit  $\lim_{x \rightarrow a} f(x)$  exists and is a finite number, hence suppose that  $\lim_{x \rightarrow a} f(x) = \ell$ . Then for every  $\varepsilon > 0$  there exists some  $\delta > 0$  such that  $|f(x) - \ell| < \varepsilon$ , for all  $x \in A$  satisfying  $0 < |x - a| < \delta$ . Hence, taking  $\varepsilon = 1$ , there exists some  $h > 0$  such that  $|f(x) - \ell| < 1$  for all  $x \in A$  with  $0 < |x - a| < h$ . By the triangle inequality this implies that  $|f(x)| < 1 + |\ell|$  for all such  $x$ . Set now  $L = 1 + |\ell|$  if  $a \notin A$ , otherwise set  $L = \max\{1 + |\ell|, f(a)\}$ . Then it is easy to see that  $|f(x)| \leq L$  for all  $x \in (a - h, a + h) \cap A$ .  $\square$

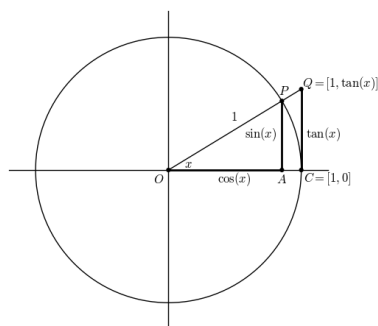
In the previous examples we often applied some basic rules, as the limit of a sum of functions is the sum of the limits, the limit of a product is the product of the limits, and the limit of a quotient is the quotient of the limits, provided that the particular limits exist and do not lead to one of the following expressions:  $0/0$ ,  $\infty/\infty$ ,  $0 \cdot \infty$ , and  $\infty - \infty$ .



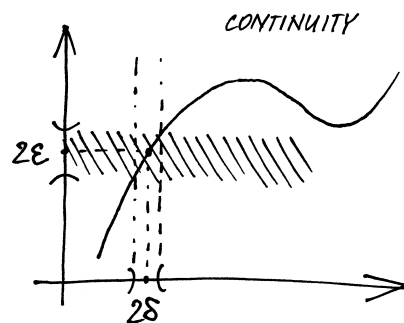
These are called “indeterminate forms”, see also the discussion in 5.2.14 and their full list includes three more types:  $0^0$ ,  $1^\infty$  and  $\infty^0$ . To address many indeterminate forms (e.g., of type  $0/0$ ,  $\infty/\infty$ ,  $0 \cdot \infty$ , etc.), we use the so-called “l’Hopital’s rule”, which we will examine later (see 5.3.10). However, let us begin with an example that demonstrates a pedestrian approach to deal with an indeterminate form of type  $0/0$ .

**5.B.31.** Show that  $\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1$ .

**Solution.** Consider the first quadrant in the unit circle  $S^1$ , and its arbitrary point  $P(x) = [\cos(x), \sin(x)]$ , where  $x$  is running in the open interval  $(0, \frac{2\pi}{2}) \subset \mathbb{R}$ . Obviously, the length of the arc joining the points  $P(x)$  and  $C := [1, 0]$ , equals to  $x$ . So we apparently have  $\sin(x) < x$  for all  $x \in (0, \frac{\pi}{2})$ , see also the figure given at the r.h.s. On the other hand, the value  $\tan(x)$  is the distance between the points  $Q(x) = [1, \tan(x)]$  and  $C$ , and we see that  $x < \tan(x)$ , for all  $x \in (0, \frac{\pi}{2})$ . Altogether,  $\sin(x) < x < \frac{\sin(x)}{\cos(x)}$  which implies that



$$\cos(x) < \frac{\sin(x)}{x} < 1,$$



Notice that for the boundary points of the interval  $A$ , the definition says that value of  $f$  equals the value of the one-sided limit there. The function is said to be *right-continuous* or *left-continuous* at such a point. Every polynomial is a continuous function on the whole  $\mathbb{R}$ , see 5.2.11(2). The Thomae function is continuous at irrational real numbers only although it has limits at all rational points as well, see 5.2.11(4).



The previous theorem 5.2.13 about limit properties implies immediately all of the following claims. The same properties are true for right-continuity or left-continuity, as is easily checked.

**5.2.17. Theorem.** Let  $f$  and  $g$  be (real or complex valued) functions defined on an interval  $A \subset \mathbb{R}$  and continuous at a point  $x_0 \in A$ . Then

- (1) the sum  $f + g$  is continuous at  $x_0$
- (2) the product  $f \cdot g$  is continuous at  $x_0$
- (3) if  $g(x_0) \neq 0$ , then the quotient  $f/g$  is well-defined on some neighbourhood of  $x_0$  and is continuous at  $x_0$ .
- (4) If a continuous function  $h$  is defined on a neighbourhood of  $f(x_0)$  of the real-valued function  $f$ , then the composite function  $h \circ f$  is defined on a neighbourhood of  $x_0$  and is continuous at  $x_0$ .

**PROOF.** Statements (1) and (2) are clear. For property (3): if  $g(x_0) \neq 0$ , then the  $\varepsilon$ -neighbourhood of the number  $g(x_0)$  does not contain zero for a sufficiently small  $\varepsilon > 0$ . By the continuity of  $g$ , it follows that on a sufficiently small  $\delta$ -neighbourhood of the point  $x_0$ ,  $g$  is non-zero and the quotient  $f/g$  is thus well-defined there. It is continuous at  $x_0$  by the previous theorem.

(4) Choose a neighbourhood  $\mathcal{O}$  of  $h(f(x_0))$ . By the continuity of  $h$ , there is a neighbourhood  $\mathcal{O}'$  of  $f(x_0)$  which is mapped into  $\mathcal{O}$  by  $h$ . The continuous function  $f$  maps some sufficiently small neighbourhood of the point  $x_0$  into the neighbourhood  $\mathcal{O}'$ . This is the definition property of continuity, so the proof is finished.  $\square$

**5.2.18.** We consider some basic relations between continuous mappings and the topology of the real numbers. They exploit the highly non-trivial characterization of compact sets in theorem 5.2.8.

for all  $x \in (0, \frac{\pi}{2})$ . Invoking now the squeeze theorem, one deduces that  $\lim_{x \rightarrow 0^+} \frac{\sin(x)}{x} = 1$ . Based then on the fact that the function  $f(x) = \sin(x)/x$  defined for all  $x \neq 0$  is even, i.e.,  $f(-x) = f(x)$ , we get

$$\lim_{x \rightarrow 0^-} \frac{\sin(x)}{x} = \lim_{x \rightarrow 0^+} \frac{\sin(x)}{x} = 1.$$

This shows that the both one-sided limits exist and have the same value, hence our claim follows. Notice that may exist other alternative ways to prove the statement.  $\square$

**5.B.32.** Based on the result from 5.B.31, show that:

(a)  $\lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x} = 0;$

(b)  $\lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x^2 \sin(x^2)} = \infty.$   $\circ$

**5.B.33.** Compute  $\lim_{x \rightarrow x_0} f(x)$  for the following cases:

(a)  $f(x) = \frac{x - 2}{\sqrt{x^2 - 4}}$ , and  $x_0 = 2$ ,

(b)  $f(x) = \frac{\sin(\sin(x))}{x}$ , and  $x_0 = 0$ ,

(c)  $f(x) = \frac{\sin^2(x)}{x}$ , and  $x_0 = 0$ .  $\circ$

**5.B.34.** Explain why the limit of the following functions as  $x \rightarrow 0$ , does not exist:

$$f(x) = \begin{cases} 1/x, & \text{if } x > 0, \\ 1, & \text{if } x \leq 0, \end{cases} \quad g(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0, \end{cases}$$

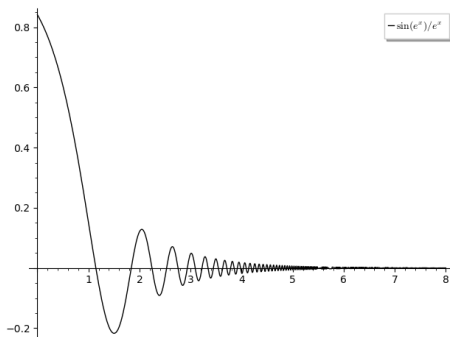
$$h(x) = \frac{|\sin(x)|}{\sin x}, \quad k(x) = \text{sign}(x). \quad \circ$$

**5.B.35.** Suppose that  $f : (0, +\infty) \rightarrow \mathbb{R}$  is a function satisfying  $|e^x f(x) - 2e^x| \leq |\sin(e^x)|$ , for any  $x \in (0, +\infty)$ . Show that  $\lim_{x \rightarrow +\infty} f(x) = 2$ .

**Solution.** The given relation  $|e^x f(x) - 2e^x| \leq |\sin(e^x)|$  can be equivalently written as

$$\begin{aligned} & -|\sin(e^x)| \leq e^x f(x) - 2e^x \leq |\sin(e^x)|, \\ \text{or} \quad & 2e^x - |\sin(e^x)| \leq e^x f(x) \leq |\sin(e^x)| + 2e^x, \\ \text{or} \quad & 2 - \left| \frac{\sin(e^x)}{e^x} \right| \leq f(x) \leq \left| \frac{\sin(e^x)}{e^x} \right| + 2. \quad (*) \end{aligned}$$

Now, it is easy to prove that  $\lim_{x \rightarrow +\infty} \frac{\sin(e^x)}{e^x} = 0$ , which also follows from the graph of this function, see here:



TOPOLOGICAL CHARACTERIZATION OF CONTINUITY

**Theorem.** Let  $f : A \subset \mathbb{R} \rightarrow \mathbb{R}$  be a function defined on an interval  $A$ . Then:

- (1)  $f$  is continuous if and only if the inverse image  $f^{-1}(U)$  of every open set  $U \subset \mathbb{R}$  is an open set in  $A$ .
- (2) If  $f$  is continuous, then the inverse image  $f^{-1}(W)$  of every closed set  $W \subset \mathbb{R}$  is a closed set in  $A$ .
- (3) If  $f$  is continuous, then the image  $f(K)$  of every compact set  $K \subset A$  is a compact set in  $A$ .
- (4) If  $f$  is continuous, then  $f$  attains both its maximum and its minimum on every compact set  $K$ .

**PROOF.** (1) Consider a point  $x_0 \in f^{-1}(U)$ . There is a neighbourhood  $\mathcal{O}$  of  $f(x_0)$  which is contained in  $U$  since  $U$  is open. Hence there is a neighbourhood  $\mathcal{O}'$  of  $x_0$  which is mapped into  $\mathcal{O}$  and thus is contained in the inverse image. Therefore, every point of the inverse image is an interior point, i.e.,  $f^{-1}(U)$  is open.

Conversely, if  $f^{-1}(U)$  is open for each open  $U$ , then taking any  $\varepsilon$ -neighbourhood of  $f(x_0)$ , its pre-image will be an open neighbourhood of  $x_0$  satisfying the condition from the definition of the continuity.

(2) Consider a limit point  $x_0$  of the inverse image  $f^{-1}(W)$  and a sequence  $x_i, f(x_i) \in W$ , which converges to  $x_0$ . From the continuity of  $f$ , it follows that  $f(x_i)$  converges to  $f(x_0)$  (cf. the convergence test 5.2.15). Since  $W$  is closed,  $f(x_0) \in W$ . Thus, all limit points of the inverse image of the set  $W$  are contained in  $f^{-1}(W)$ .

(3) Choose any open cover  $U_\alpha$  of  $f(K)$ . The inverse images of all  $U_\alpha$  are open and thus create an open cover of the set  $K$ . Select a finite subcover from it. Then finitely many of the corresponding images cover the original set  $f(K)$ .

(4) Since the image of a compact set is again a compact set, the image must be bounded and it contains both the supremum and the infimum. Hence it follows that these must also be the maximum and the minimum, respectively.  $\square$

Notice, a complex valued function is continuous if and only if its real and imaginary components are continuous. The first three claims of the theorem remain valid with very slight modification of the proof. (Check it yourselves!)

**5.2.19.** There are two very useful consequences of the previous theorem.

For instance, via the squeeze theorem one can show that

$$\lim_{y \rightarrow +\infty} \frac{\sin y}{y} = 0$$

and then the replacement  $y = e^x$  yields our claim, or directly, via Sage type:  $\lim(\sin(e^x)/e^x, x = +\infty)$ . Hence, we also deduce that

$$\lim_{x \rightarrow +\infty} \left( 2 - \left| \frac{\sin(e^x)}{e^x} \right| \right) = 2 = \lim_{x \rightarrow +\infty} \left( \left| \frac{\sin(e^x)}{e^x} \right| + 2 \right)$$

and by (\*) we are done via the squeeze theorem.  $\square$

**5.B.36.** For some  $\alpha > 0$  let  $g$  the function defined by

$$g(x) = \lim_{\alpha \rightarrow +\infty} \left( \frac{(2x-3)\alpha^2 + 4x\alpha + 2}{x+\alpha} \cdot \sin\left(\frac{1}{\alpha}\right) \right).$$

Show that  $g$  is a line which is a median of the triangle  $ABC$  with corners the points  $A = [1, -1]$ ,  $B = [6, 2]$ ,  $C = [-2, 0]$ .  $\circ$

*So far we tried to convince the reader to use common sense when reasoning about the limit behaviour of functions and prevent any mindless following of rules. On the other hand, beyond the l'Hopital's rule which is based on the notion of derivatives and we will discuss later (see also 5.3.10), there are diverse useful "rules" or "tricks" to deal with the indeterminate expressions, which are not necessarily based on derivatives. Let us mention some of them.*

a) Recall that if  $f(x) > 0$  for all  $x$  in the domain of  $f$ , then

$$f(x)^{g(x)} = e^{\ln(f(x)^{g(x)})} = e^{g(x) \cdot \ln(f(x))}.$$

Since the function  $e^x$  is continuous and injective on  $\mathbb{R}$ , we get

$$\lim_{x \rightarrow x_0} f(x)^{g(x)} = e^{\lim_{x \rightarrow x_0} (g(x) \ln(f(x)))} = e^{\lim_{x \rightarrow x_0} g(x) \cdot \lim_{x \rightarrow x_0} \ln(f(x))}.$$

Moreover, it is easy to prove that

$$\lim_{x \rightarrow x_0} (g(x) \cdot \ln f(x)) = a \in \mathbb{R} \implies \lim_{x \rightarrow x_0} f(x)^{g(x)} = e^a,$$

$$\lim_{x \rightarrow x_0} (g(x) \cdot \ln f(x)) = +\infty \implies \lim_{x \rightarrow x_0} f(x)^{g(x)} = +\infty,$$

$$\lim_{x \rightarrow x_0} (g(x) \cdot \ln f(x)) = -\infty \implies \lim_{x \rightarrow x_0} f(x)^{g(x)} = 0.$$

b) Some further useful rules to remember are here:

$$\lim_{x \rightarrow +\infty} \frac{c}{x^\alpha} = 0, \quad \lim_{x \rightarrow +\infty} \frac{x^\alpha}{x^\beta} = 0, \quad \lim_{x \rightarrow +\infty} \frac{x^\beta}{a^x} = 0,$$

$$\lim_{x \rightarrow +\infty} \frac{a^x}{b^x} = 0, \quad c \in \mathbb{R}, \quad 0 < \alpha < \beta, \quad 1 < a < b.$$

You may try to verify them (a way to determine the third limit is based on the l'Hopital's rule, for example).

**5.B.37.** Compute the following limits:

(a)  $\lim_{x \rightarrow +\infty} \frac{3^{x+1} + x^5 - 4x}{3^x + 2x + x^2},$

(b)  $\lim_{x \rightarrow +\infty} \frac{4^x - 8x^6 - 2^x - 167}{3^x - 45x - \sqrt{11}\pi^{x+12}},$

**Corollary.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be continuous. Then

- (1) the image of every interval is again an interval,
- (2)  $f$  takes all the values between the maximal and the minimal one on the closed interval  $[a, b]$ .



**PROOF.** (1) Consider an open interval  $A$  and suppose there is a point  $y \in \mathbb{R}$  such that  $f(A)$  contains points less than  $y$  as well as points greater than  $y$ , but  $y \notin f(A)$ . Put  $B_1 = (-\infty, y)$  and  $B_2 = (y, \infty)$ . These are open sets, and the union of their inverse images  $A_1 = f^{-1}(B_1) \subset A$  and  $A_2 = f^{-1}(B_2) \subset A$  contains  $A$ .  $A_1$  and  $A_2$  are open, disjoint, and they both have a non-empty intersection with  $A$ . Thus there is a point  $x \in A$  which does not lie in  $A_1$  but is a limit point of  $A_1$ . It is in  $A_2$ , which is impossible for two disjoint open sets.

Thus it is proved that if there is a point  $y$  which does not belong to the image of the interval, then either all of the values must be above  $y$  or they all must be below. It follows that the image is again an interval. Notice that the boundary points of this interval may or may not lie in the image.

If the domain interval  $A$  contains one of its boundary points, then the continuous function must map it to a limit point or an interior point of the image of the interior of  $A$ . This verifies the statement.

(2) This statement immediately follows from the previous one (and the above theorem) since the image of a closed bounded interval (i.e. a compact set) is again a closed interval.  $\square$

**5.2.20.** We conclude this introductory discussion by two more theorems which provide useful tools for calculating limits. Notice that we assume that functions are defined on all of  $\mathbb{R}$ . Actually we are only interested in  $f$  on a neighbourhood of one point  $a$ , while  $g$  has to be defined on a neighbourhood of one point  $b$  only.

#### LIMITS OF COMPOSITE FUNCTIONS

**Theorem.** Let  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  be two functions and  $\lim_{x \rightarrow a} f(x) = b$ .

(1) If the function  $g$  is continuous at the point  $b$ , then

$$\lim_{x \rightarrow a} g(f(x)) = g\left(\lim_{x \rightarrow a} f(x)\right) = g(b).$$

(2) If the limit  $\lim_{y \rightarrow b} g(y)$  exists and  $f(x) \neq b$  holds for all  $x \neq a$  from some neighbourhood of the point  $a$ , then

$$\lim_{x \rightarrow a} g(f(x)) = \lim_{y \rightarrow b} g(y).$$

<sup>9</sup>This result is usually called the Weierstrass theorem, but it is also known (especially in Czech literature) as the Bolzano's theorem. Bernard Bolzano apparently used such result as a technical lemma when proving his Bolzano-Weierstrass theorem mentioned earlier.

(c)  $\lim_{x \rightarrow 0} \frac{\sqrt{1+x} - \sqrt{1-x}}{x}$ ,

(d)  $\lim_{x \rightarrow \pi/4} \frac{\cos x - \sin x}{\cos(2x)}$ . ○

**5.B.38.** Consider a polynomial function  $f : \mathbb{R} \rightarrow \mathbb{R}$  whose degree is determined by the result of rolling a die. Find the probability  $P(A)$ , where  $A$  denotes the following event:

$$\lim_{x \rightarrow +\infty} \frac{x^4 + x^2 + 1}{f(x)} = 0.$$

**Solution.** We have  $f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ , where by assumption  $n \in \{1, 2, \dots, 6\}$  (and so  $f(x) \neq 0$ ). Thus one computes

$$\begin{aligned} \lim_{x \rightarrow +\infty} \frac{x^4 + x^2 + 1}{f(x)} &= \\ &= \lim_{x \rightarrow +\infty} \frac{x^4 + x^2 + 1}{a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0} \\ &= \lim_{x \rightarrow +\infty} \frac{x^4}{a_n x^n}. \end{aligned}$$

Hence we seek for the probability of having  $\lim_{x \rightarrow +\infty} \frac{x^4}{a_n x^n} = 0$ . It is now reasonable to consider the following cases:

1) For  $n = 1, 2, 3$  we get  $\lim_{x \rightarrow +\infty} \frac{x^4}{a_n x^n} = \pm\infty$ .

2) For  $n = 4$  we get  $\lim_{x \rightarrow +\infty} \frac{x^4}{a_4 x^4} = \frac{1}{a_4} \neq 0$ .

3) For  $n = 5, 6$  we obtain  $\lim_{x \rightarrow +\infty} \frac{x^4}{a_n x^n} = 0$ .

Therefore, the favourable cases are those with  $n = 5, 6$ . Since the sample space is  $\Omega = \{1, 2, \dots, 6\}$  we deduce that the probability in question equals to  $P(A) = \frac{2}{6} = \frac{1}{3}$ . □

**5.B.39.** Consider the functions  $f(x) = e^{4x+2}$  and  $g(x) = \ln(x^2)$ , defined on  $\mathbb{R} \setminus \{0\}$  and  $[1, e^8]$ , respectively.

(a) Show the composition  $\varphi = g \circ f$  is determined, and present its explicit form. Confirm your result via Sage and the command `compose`.

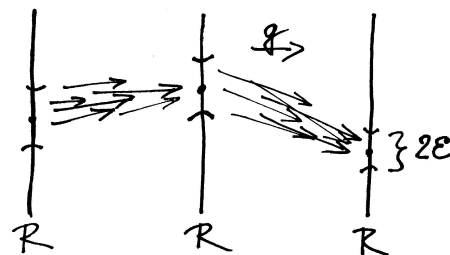
(b) Evaluate the limit  $L = \lim_{x \rightarrow 0} \frac{\varphi(x) - \sin^3(x) - 4}{x}$ .

**Solution.** (a) Let us denote by  $A = \mathbb{R} \setminus \{0\}$  and  $B = [1, e^8]$  the two given domains. The composition  $\varphi(x) = g(f(x))$  is determined, if the set  $X := \{x \in A : f(x) \in B\}$  is non empty, i.e.,  $X \neq \emptyset$ . We compute

$$\begin{aligned} X &= \{x \in \mathbb{R}^* : 1 \leq e^{4x+2} \leq e^8\} \\ &= \{x \in \mathbb{R}^* : e^0 \leq e^{4x+2} \leq e^8\} \\ &= \{x \in \mathbb{R}^* : 0 \leq 4x + 2 \leq 8\} \\ &= \left\{x \in \mathbb{R}^* : -\frac{1}{2} \leq x \leq \frac{3}{2}\right\} = \left[-\frac{1}{2}, 0\right) \cup \left(0, \frac{3}{2}\right], \end{aligned}$$

where here  $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ . Therefore, based on the basic properties of logarithms, we get  $\varphi : X \rightarrow \mathbb{R}$  with type

$$\varphi(x) = g(f(x)) = \ln(f(x)^2) = \ln e^{2(4x+2)} = 2(4x + 2).$$



**PROOF.** The first proposition can be proved similarly to 5.2.17(4). From the continuity of  $g$  at the point  $b$ , it follows that for any neighbourhood  $V$  of the value  $g(b)$ , we can find a sufficiently small neighbourhood  $U$  of the point  $b$  whose values of  $g$  lies in  $V$ . However, if  $f$  has limit  $b$  at the point  $a$ , then  $f$  will hit  $U$  by all its values  $f(x)$  for  $x \neq a$  from some sufficiently small neighbourhood of the point  $a$ , which verifies the first statement.

(2) Even if we cannot use the continuity of  $g$  at the point  $b$ , the previous reasoning will be true if we ensure that all points  $x \neq a$  from sufficiently small neighbourhoods of  $a$  are mapped into a neighbourhood of  $b$  by the function  $f$ , but  $f(x) \neq b$  for all such points. □

**5.2.21. Who or what is in the ZOO.** We have begun to



build a menagerie of functions with polynomials and functions which can be created from them “piece-wise”. Moreover, we have derived many properties for a huge class of continuous functions. However, except for polynomials, we do not have many practically manageable examples at our disposal. We consider the quotients of polynomials.

Let  $f$  and  $g$  be two polynomials in the real variable  $x$  with complex coefficients  $a_i \in \mathbb{C}$ .

The function  $h : \mathbb{R} \setminus \{x \in \mathbb{R}, g(x) = 0\} \rightarrow \mathbb{C}$ ,

$$h(x) = \frac{f(x)}{g(x)}$$

is well-defined for all real  $x$  except for the roots of the polynomial  $g$ . Such functions are called *rational functions*. From theorem 5.2.17, it follows that rational functions are continuous at all points of their domains. At the points where the denominators of real rational functions vanish, they can have

- a finite limit,
- an infinite limit, supposing the one-sided infinite limits are equal,
- different one-sided infinite limits.

For the case of a finite limit, it is necessary that the point is a common root of both  $f$  and  $g$  and that its multiplicity in  $f$  is at least as large as in  $g$ . Then the domain of the rational function can be extended by this point, defining it to take the value of the limit there. Then the new function is continuous at this point, too.

The possibilities are illustrated in the diagram showing the values of the function

$$h(x) = \frac{(x - 0.05a)(x - 2 - 0.2a)(x - 5)}{x(x - 2)(x - 4)}$$

To verify this very last computation in Sage, use the cell

```
f(x)=e**(4*x+2); ln(f(x)^2).simplify_full()
```

which returns  $8 * x + 4$ . The same returns the method suggested in the statement, via the command `compose`, which we implement by the cell

```
f(x)=e^(4*x+2); g(x)=ln(x^2)
h = compose(g, f)
show(h(x).simplify_full())
```

(b) The explicit form of  $\varphi$  is known by (a), which makes it easy to compute the limit at hand:

$$L = \lim_{x \rightarrow 0} \frac{8x - \sin^3(x)}{x} = \lim_{x \rightarrow 0} \left( 8 - \sin^2(x) \cdot \frac{\sin(x)}{x} \right)$$

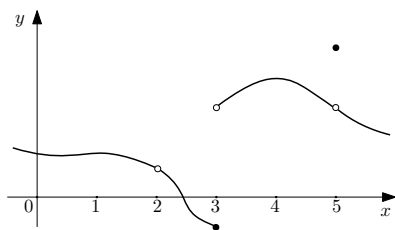
$$= 8 - \lim_{x \rightarrow 0} \sin^2(x) \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 8 - 0 \cdot 1 = 8.$$

□



We have exploited the continuity of several functions in an implicit or explicit way many times already. Now it is time to return back to the essence and play a little bit with the concept itself.

**5.B.40.** The figure below shows the graph of a function  $y = f(x)$ . Indicate the points where  $f$  is discontinuous and explain why.

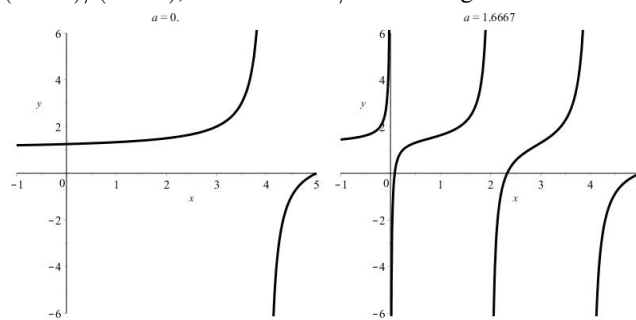


**Solution.** There is a discontinuity at  $x = 2$ , since there  $f$  is not defined (the graph has a break that is represented by the small circle). Another discontinuity appears at  $x = 3$  (a jump discontinuity). Indeed, although  $f(3)$  is defined (is a negative number), we see that the left and right limits are different, so  $\lim_{x \rightarrow 3} f(x)$  does not exist. Finally, according to the figure at  $x = 5$  the value  $f(5)$  is defined and moreover  $\lim_{x \rightarrow 5} f(x)$  exists. However, we see that  $\lim_{x \rightarrow 5} f(x) \neq f(5)$ . Thus  $f$  is discontinuous also at  $x = 5$  (we may refer to such a discontinuity by the term removable discontinuity). □

**5.B.41.** Sketch the graph of a function which is continuous everywhere, except of a removable discontinuity at  $x = 1$ , a jump discontinuity at  $x = 3$ , and a discontinuity at  $x = 4$  which is not of the two previous types.

**Solution.** There are of course many such functions. An example is given here:

for  $a = 0$  on the left (thus it is the simpler rational function  $(x - 5)/(x - 4)$ ) and for  $a = 5/3$  on the right.



Notice that the situation gets more complicated for the complex rational functions. Their real and imaginary parts are again (real) rational functions (see the exercise ??) and the above options are true for only each of the components separately.

**5.2.22. Power functions and the exponential.**

We have met the simple power functions  $x \mapsto x^n$  with natural exponents  $n = 0, 1, 2, \dots$  when building the polynomials. The meaning of the function  $x \mapsto x^{-1}$ , defined for all  $x \neq 0$ , is also clear.

Now, we extend this definition to a general power function  $x^y$  with an arbitrary (constant)  $y \in \mathbb{R}$  and  $x > 0$ .

Changing our mind about constants and variables, we obtain the values of exponential function  $x^y$  with arbitrary (constant)  $x > 0$  called the base.

With natural  $y$  and  $z$ , we know the following rule

(1)  $x^{y+z} = x^y \cdot x^z$

and its consequence

(2)  $x^{y \cdot z} = (x^y)^z.$

We shall extend this definition to all positive real  $x > 0$  and all real  $y$ .

EXPONENTIAL AND POWER FUNCTIONS

**Theorem.** There is the unique function  $f(y) = x^y$ , defined for all  $x > 0, y \in \mathbb{R}$ , separately continuous in both variables  $x$  and  $y$  (i.e. we consider the other one as a constant when checking the continuity) and satisfying (1),  $f(0) = 1$ , and  $f(1) = x$ . This function also satisfies (2).

We call this function the exponential function with the base  $x$ , or the power function with exponent  $y$ .

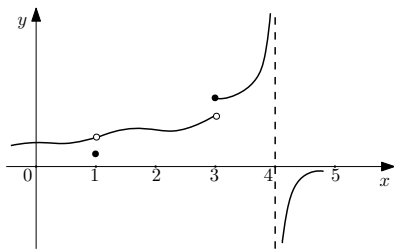
**PROOF.** Of course, we want to extend the well known values of the powers  $x^n$  with  $x$  rational and  $n$  integral, which are also direct consequences of (1) (and then automatically satisfy (2)).

For a negative integer  $-a$ , we have to define (in view of (1))

$$x^{-a} = (x^a)^{-1} = (x^{-1})^a.$$

Further, we want the equality  $b^n = x$  for  $n \in \mathbb{N}$  to imply that  $b$  is the  $n$ -th root of  $x$ , i.e.  $b = x^{1/n}$  (again a consequence of





Notice the discontinuity at  $x = 4$ : Can you analyze why this differs from the previous two?  $\square$

**5.B.42.** Explain why the functions  $f, g : (0, +\infty) \rightarrow \mathbb{R}$  defined by  $f(x) = x^x$  and  $g(x) = x^{\cos(x)}$ , respectively, are continuous.  $\circ$

**5.B.43.** If  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  are continuous functions with  $f(2) = 6$  and  $\lim_{x \rightarrow 2} (4f(x) - g(x)) = 2$ , find  $g(2)$ .  $\circ$

**5.B.44.** Present an example of a function defined on  $\mathbb{R}$  which is everywhere continuous except at  $x = 0$ .  $\circ$

**5.B.45.** Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$

$$f(x) = \begin{cases} x, & \text{if } x \in \mathbb{Q}, \\ -x, & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

Determine the points  $x_0 \in \mathbb{R}$ , if any, where  $f$  is continuous.

**Solution.** The given function  $f$  is continuous only at  $x_0 = 0$ . For a confirmation, consider any sequence  $(x_n)$  or  $\mathbb{R}$  with  $x_n \rightarrow 0$  and  $x_n \neq 0$  for all  $n$ . Then obviously  $(-x_n) \rightarrow 0$  as well and hence  $f(x_n) \rightarrow 0$ . The claim now follows by combining the definition of continuity of a given function at a chosen point with Proposition 5.2.15. On the other hand, for any non-zero real number  $a$  we can find sequences  $(x_n)$  and  $(y_n)$  with  $x_n \rightarrow a$  and  $y_n \rightarrow a$ , which however satisfy  $f(x_n) \rightarrow a$  and  $f(y_n) \rightarrow -a$ , respectively. This implies that  $f$  is discontinuous at any  $a \in \mathbb{R} \setminus \{0\}$ .  $\square$

**5.B.46.** Determine the interval(s) on which each of the following functions is continuous:

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0, \end{cases}$$

$$g(x) = \begin{cases} x \sin\left(\frac{1}{x}\right), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

**5.B.47.** Consider the function

$$g(x) = \begin{cases} \frac{\tan(x)}{\pi - 2x} + x, & \text{if } x \in \left(0, \frac{\pi}{2}\right), \\ \frac{2x + 1}{2}, & \text{if } x \in \left[\frac{\pi}{2}, +\infty\right). \end{cases}$$

Show that  $g(x)$  is continuous at  $x_0$ , where  $x_0$  is the value of the determinant  $|A|$ , and  $A \in \text{Mat}_4(\mathbb{R})$  is a real  $4 \times 4$  invertible matrix satisfying  $A = (1/|A|)B$ , for some square matrix  $B \in \text{Mat}_4(\mathbb{R})$  with  $\det(B) = |B| = (\pi/2)^5$ .

(1) since this requests  $(x^{\frac{1}{n}})^n = x^{n \cdot \frac{1}{n}} = x$ . We verify that such  $b$ 's always exist for positive real numbers  $x$ .

By factoring out  $y_2 - y_1$  in  $y_2^n - y_1^n$ , or otherwise, we see that the function  $y \mapsto y^n$  is strictly increasing for  $y > 0$ . Choose  $x > 0$  and consider the set  $B = \{y \in \mathbb{R}, y > 0, y^n \leq x\}$ . This is a non-empty set bounded from above, so set  $b = \sup B$ . A power function with natural exponent  $n$  is continuous, thus  $b^n = x$ . Indeed, surely  $b^n \leq x$ . If the inequality is strict, then there is a number  $y$  such that  $b^n < y^n < x$ , which implies that  $b < y$ . This contradicts the definition of  $b$  as a supremum.

Thus the power function is suitably defined for all rational numbers  $a = \frac{p}{q}$ . For  $x > 0$ , we set  $x^a = (x^p)^{\frac{1}{q}} = (x^{\frac{1}{q}})^p$ .

Finally, we notice that for the values  $0 < a \in \mathbb{Q}$  and fixed  $x > 1$ ,  $x^a$  is strictly increasing for rational  $a$ 's. Therefore, for general  $0 < a \in \mathbb{R}$  and  $1 < x$  we can define

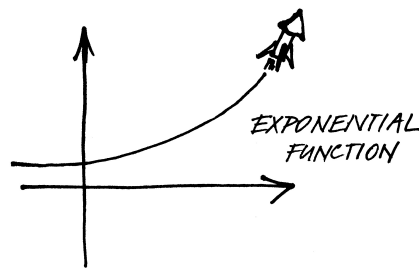
$$x^a = \sup\{x^y, y \in \mathbb{Q}, y \leq a\}.$$

As before,  $x^{-a} = \frac{1}{x^a}$ .

For  $0 < x < 1$ , proceed analogously with care for the inequality signs, or set  $x^a = \left(\frac{1}{x}\right)^{-a}$ . For  $x = 1$ , define  $1^a = 1$  for any  $a$ , while  $0^a = 0$ .

Now, the power function  $x \mapsto x^a$  is defined for all  $x \in [0, \infty)$  and  $a \in \mathbb{R}$ . Notice, the requested continuity in both parameters fixed all our choices above. In particular, our function is continuous also in the parameter  $a$  and we have constructed also the exponential functions  $c^y$  for constants  $c > 0$ .

The property (1) used when defining the power function  $x^a$  for integral  $a$  implied also (2), and they both manifestly survived through the construction for any choices of the values  $x, y$ .  $\square$



**5.2.23. Logarithmic functions.** The exponential function  $f(x) = a^x$  is increasing for  $a > 1$  and decreasing for  $0 < a < 1$ . Thus in both cases, there is a function  $f^{-1}(x)$  inverse to it. This function is called the *logarithmic function with base  $a$* .



We write  $\ln_a(x)$  and the defining property is

$$\ln_a(a^x) = x.$$

The equalities 5.2.22(1),(2) are thus equivalent to

$$\ln_a(x \cdot y) = \ln_a(x) + \ln_a(y), \quad \ln_a(x^y) = y \cdot \ln_a(x).$$

Logarithmic functions are defined only for positive arguments and they are, on the entire domain, increasing if  $a > 1$  and decreasing for  $0 < a < 1$ . Moreover,  $\ln_a(1) = 0$  holds for every  $a$ .

**Solution.** By assumption,  $A$  and  $B$  are real  $4 \times 4$  matrices satisfying the relation  $B = |A|A$ . In Chapter 2 we saw that  $\det(\lambda C) = \lambda^n \det(C)$  for any real  $n \times n$ -matrix  $C$  and  $\lambda \in \mathbb{R}$ , see ???. Thus we get  $|B| = \left| |A|A \right| = |A|^4 |A| = |A|^5$ , that is,  $(\pi/2)^5 = |A|^5$  or equivalently  $|A| = \pi/2$ . Hence  $x_0 = \pi/2$  and we see that

$$\begin{aligned} \lim_{x \rightarrow \frac{\pi}{2}^+} f(x) &= \lim_{x \rightarrow \frac{\pi}{2}^+} \left( x + \frac{1}{2} \right) = \frac{\pi}{2} + \frac{1}{2}, \\ \lim_{x \rightarrow \frac{\pi}{2}^-} f(x) &= \lim_{x \rightarrow \frac{\pi}{2}^-} \frac{\tan x}{\pi - 2x} + \lim_{x \rightarrow \frac{\pi}{2}^-} x \\ &= \frac{1}{2} \lim_{x \rightarrow \frac{\pi}{2}^-} \frac{\tan(\frac{\pi}{2} - x)}{(\frac{\pi}{2} - x)} + \frac{\pi}{2} = \frac{\pi}{2} + \frac{1}{2}. \end{aligned}$$

Here we have used the trigonometric identity  $\tan(\frac{\pi}{2} - x) = \tan(x)$  (try to prove this), and moreover that

$$\lim_{y \rightarrow 0} \frac{\tan y}{y} = 0, \quad (*)$$

where above one should replace  $\frac{\pi}{2} - x$  by  $y$  and  $x \rightarrow \frac{\pi}{2}^-$  with  $y \rightarrow 0$ . Therefore, the one-sided limits coincide and since they equal to  $f(\frac{\pi}{2}) = \frac{\pi}{2} + \frac{1}{2}$ ,  $f$  is continuous at  $x_0$ . To be complete, let us prove (\*). For this one uses the result from 5.B.31 and that the limit of a product is the product of the limits, if both limits are defined:

$$\begin{aligned} \lim_{y \rightarrow 0} \frac{\tan y}{y} &= \lim_{y \rightarrow 0} \frac{\sin y}{\cos y} = \lim_{y \rightarrow 0} \frac{\sin y}{y} \cdot \lim_{y \rightarrow 0} \frac{1}{\cos y} \\ &= 1 \cdot \frac{1}{\cos 0} = 1. \end{aligned}$$

□

Given a real number  $x \in \mathbb{R}$  its floor (also called the integer part of  $x$ ), denoted by  $\lfloor x \rfloor$  (or by  $[x]$ ), is the greatest integer which is less than or equal to  $x$ , i.e.,  $\lfloor x \rfloor = \max\{a \in \mathbb{Z} : a \leq x\}$ . Equivalently, we can view  $\lfloor x \rfloor$  as the unique integer satisfying  $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$ . Hence for example we have  $\lfloor x \rfloor = 1$  for  $1 \leq x < 2$ ,  $\lfloor x \rfloor = 2$  for  $2 \leq x < 3$ , etc. The floor function or greatest integer function is then defined by  $f(x) = \lfloor x \rfloor \in \mathbb{Z}$ , for all  $x \in \mathbb{R}$ . Obviously, for any integer  $n \in \mathbb{Z}$  we have  $f(n) = \lfloor n \rfloor = n$ .

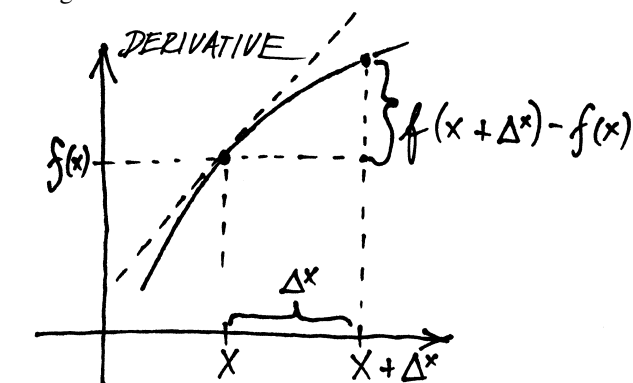
In the same vein one can define the “ceiling function” (also known as the “least integer function”)  $x \mapsto \lceil x \rceil$ , where  $\lceil x \rceil$  is the smallest integer  $\geq x$ , called the “ceiling of  $x$ ”, i.e.,  $\lceil x \rceil = \min\{a \in \mathbb{Z} : a \geq x\}$ . In addition,  $\{x\} = x - \lfloor x \rfloor$  is called the “fractional part of  $x$ ”, and the assignment  $x \mapsto \{x\}$  induces the so called “fractional part function”. Obviously,  $0 \leq \{x\} < 1$  and for nonnegative real numbers, the fractional part is just the part of the number after the decimal (but this is not the case for negative real numbers).

Nevertheless, together with the floor function for which we devote some space in 5.B.48, these functions are some of the simplest examples of discontinuous functions, and they can be quite useful in constructing more wild examples. Moreover, they have all important applications in integral

There is an extremely important value of  $a$ , namely the number  $e$ , see the paragraph 5.4.1, also known as Euler’s number. The function  $\ln_e(x)$  is called the *natural logarithm* and is denoted by  $\ln(x)$  (i.e., omitting the base  $e$ ).

### 3. Derivatives

When talking about polynomials, the rate at which the function changes at a given point of its domain was already discussed (see the paragraph 5.1.6). It is the quotient 5.1.6(1), which expresses the slope of the secant line between the points  $[x, f(x)] \in \mathbb{R}^2$  and  $[x + \delta x, f(x + \delta x)] \in \mathbb{R}^2$  for a (small) change  $\delta x$  of the input variable. This reasoning is correct for any real or complex function  $f$ . It is only necessary to work with the concept of the limit, instead of the intuitive “small change”  $\delta x$ .



DERIVATIVE OF A FUNCTION OF A REAL VARIABLE

**5.3.1. Definition.** Let  $f$  be a real or complex function defined on an interval  $A \subset \mathbb{R}$  and  $x_0 \in A$ . If the limit

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = a$$

exists, the function  $f$  is said to be differentiable at  $x_0$ , provided  $a$  is finite. The value of the derivative at  $x_0$ , namely  $a$ , is denoted by  $f'(x_0)$  or  $\frac{df}{dx}(x_0)$  or  $\frac{d}{dx}f(x_0)$ .

If  $a$  is finite, the derivative is also sometimes called *proper*. If  $a$  is infinite, it is *improper*.

If  $x_0$  is one of the boundary points of  $A$ , we arrive at *one-sided derivatives* (i.e., left-sided derivative and right-sided derivative).

If a function has a derivative at  $x_0$ , the function is said to be *differentiable* at  $x_0$ . A function which is differentiable at every point of a given interval is said to be differentiable on the interval.

Obviously, the derivative of a complex valued function  $(f(x) + i g(x))$  exists if and only if the derivatives of both real and imaginary parts  $f$  and  $g$  exist (see the elementary properties of limits). Then

$$(f(x) + i g(x))' = f'(x) + i g'(x).$$



calculus, in number theory and in computer science. Hence we will meet them again in the next chapter and in Chapter 11, the latter devoted to number theory. For more details on the functions  $\lceil x \rceil$  and  $\lfloor x \rfloor$  we refer to the problem 5.E.57 in the final section.

**5.B.48. The floor function.** Let  $f(x) = \lfloor x \rfloor$  be the floor function.

- (1) Compute  $f(-0.4)$ ,  $f(0.4)$ ,  $f(0.99)$ ,  $f(3/2)$ ,  $f(\sqrt{7})$ ,  $f(3)$ ,  $f(\pi)$ ,  $f(9/2)$  and  $f(19/2)$ .
- (2) Show that  $f(x+n) = f(x) + n$ , for all  $x \in \mathbb{R}$  and  $n \in \mathbb{Z}$ . Next present an example confirming this relation.
- (3) Show that the relation  $f(nx) = nf(x)$ , with  $x \in \mathbb{R}$  and  $n \in \mathbb{Z}$  is false in general.
- (4) Deduce that  $\lim_{x \rightarrow 2^+} f(x) \neq \lim_{x \rightarrow 2^-} f(x)$ . More in general, prove that  $f$  is discontinuous at any integer  $n \in \mathbb{Z}$ .
- (5) Use Sage to sketch the graph of the floor function for  $-5 \leq x \leq 5$ .

**Solution.** (1) Obviously,  $\lfloor -0.4 \rfloor = -1$ ,  $\lfloor 0.4 \rfloor = 0 = \lfloor 0.99 \rfloor$ ,  $\lfloor 3/2 \rfloor = 1$ ,  $\lfloor \sqrt{7} \rfloor = 2$ ,  $\lfloor 3 \rfloor = 3 = \lfloor \pi \rfloor$ ,  $\lfloor 9/2 \rfloor = 4$ , and  $\lfloor 19/2 \rfloor = 9$ . Notice that in Sage the floor function corresponds to the command `floor`. Hence, to verify some of our previous computations, one may type (and similarly for the cases omitted).

```
print(floor(-0.4)); print(floor(sqrt(7)))
print(floor(pi)); print(floor(19/2))
```

- (2) We need to prove that  $\lfloor x+n \rfloor = \lfloor x \rfloor + n$  for all  $x \in \mathbb{R}$  and  $n \in \mathbb{Z}$ . By definition, we have  $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$  and by adding  $n$  to all the parts we get  $\lfloor x \rfloor + n \leq x+n < \lfloor x \rfloor + n + 1$ . Recall however that for all  $x \in \mathbb{R}$  and  $m \in \mathbb{Z}$  the relations  $\lfloor x \rfloor = m$  and  $m \leq x < m + 1$  are equivalent. Thus for  $m = \lfloor x \rfloor + n$  the above inequality means  $m \leq x < m + 1$ , and hence we get  $\lfloor x+n \rfloor = m$ . Comparing now the two equalities involving  $m$ , the result follows. As an example, we have  $\lfloor 9/2 \rfloor = \lfloor 0.5 + 4 \rfloor = \lfloor 0.5 \rfloor + 4 = 0 + 4 = 4$ .
- (3) As a counterexample notice that  $\lfloor 3 \cdot \frac{1}{3} \rfloor = \lfloor 1 \rfloor = 1$ , but  $3 \cdot \lfloor \frac{1}{3} \rfloor = 3 \cdot 0 = 0$ .
- (4) By definition, we have  $\lfloor x \rfloor = 2$  for  $2 \leq x < 3$ . Thus  $\lim_{x \rightarrow 2^+} \lfloor x \rfloor = \lim_{x \rightarrow 2^+} 2 = 2$ . Similarly, since  $\lfloor x \rfloor = 1$  for  $1 \leq x < 2$  we have  $\lim_{x \rightarrow 2^-} \lfloor x \rfloor = \lim_{x \rightarrow 2^-} 1 = 1$ . This shows that  $f$  is discontinuous at  $x = 2$ , and similarly at any integer  $n \in \mathbb{Z}$ , i.e.,  $\lim_{x \rightarrow n^+} \lfloor x \rfloor = n = f(n) \neq \lim_{x \rightarrow n^-} \lfloor x \rfloor = n - 1$ .
- (5) Let us now present the graph of the floor function. For convenience, we will use small circles to indicate the jump discontinuities of  $f$  at the integers lying in  $[-5, 5]$ . We can do this manually in Sage by combining first the commands `floor` and `plot`, to sketch  $f(x) = \lfloor x \rfloor$ , and next the commands `point` and `circle` to draw the discontinuity points, etc. This is encoded by the following cell:

Derivatives are handled rather easily, but it takes time to derive the proper formulae for derivatives of the elementary functions in our zoo. Therefore, we present a table of derivatives of several such functions in advance. In the last column, there are references to the corresponding paragraphs where the results are proved. Notice that even though we are unable to express inverse functions to some of our functions by elementary means, we can calculate their derivatives; see 5.3.6.

DERIVATIVES OF SOME FUNCTIONS

function	domain	derivative	
polynomials $f(x)$	whole $\mathbb{R}$	$f'(x)$ is again a polynomial	5.1.6
cubic splines $h(x)$	whole $\mathbb{R}$	continuous second derivatives	5.1.9
rational functions $f(x)/g(x)$	whole $\mathbb{R}$ , except for roots of $g$	rational functions: $\frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$	5.3.5
power functions $f(x) = x^a$	interval $(0, \infty)$	$f'(x) = ax^{a-1}$	5.3.7
exponential functions $f(x) = a^x$ , $a > 0, a \neq 1$	whole $\mathbb{R}$	$f'(x) = \ln(a)a^x$	5.3.7
logarithmic function $f(x) = \ln_a(x)$ , $a > 0, a \neq 1$	interval $(0, \infty)$	$f'(x) = (\ln(a))^{-1} \frac{1}{x}$	5.3.7

The initial idea of the definition suggests that  $f'(x_0)$  allows an approximation to the function  $f$  by the straight line

$$y = f(x_0) + f'(x_0)(x - x_0).$$

This is the meaning of the following lemma, which says that replacing the constant coefficient  $f'(x_0)$  in the line's equation with a certain continuous function gives exactly the values of  $f$ . The difference between  $\psi(x)$  and  $\psi(x_0)$  on a neighbourhood of  $x_0$  then says how much the slopes of the secant lines and the tangent line at  $x_0$  differ.



**Lemma.** A real or complex function  $f(x)$  has a finite derivative at  $x_0$  if and only if there is a neighbourhood  $\mathcal{O}$  of  $x_0$  and a function  $\psi$  defined on this neighbourhood which is continuous at  $x_0$  and such that for all  $x \in \mathcal{O}$ ,

$$f(x) = f(x_0) + \psi(x)(x - x_0).$$

Furthermore,  $\psi(x_0) = f'(x_0)$ , and  $f$  is continuous at the point  $x_0$ .

**PROOF.** If  $\psi$  exists, it is of the form

$$\psi(x) = \frac{f(x) - f(x_0)}{x - x_0}$$

for all  $x \in \mathcal{O} \setminus \{x_0\}$ .

```
f=floor(x)
p=plot(f, x, -5, 5, color="steelblue",
ticks=[1,1],
legend_label=r"$\text{floor function}$")
```

Suppose  $f'(x_0)$  is the proper derivative. Then we can define the value at the point  $x_0$  as  $\psi(x_0) = f'(x_0)$ . Certainly,

$$\lim_{x \rightarrow x_0} \psi(x) = f'(x_0) = \psi(x_0).$$

Thus  $\psi$  is continuous at  $x_0$  as desired.

On the other hand if such a function  $\psi$  exists, the same procedure calculates its limit at  $x_0$ . Thus the derivative  $f'(x_0)$  exists as well and equals  $\psi(x_0)$ .

The function  $f$  is expressed in terms of the sum and product of functions continuous at  $x_0$ . Thus  $f$  is continuous at  $x_0$ .  $\square$

### 5.3.2. Geometrical meaning of the derivative.



The previous lemma leads to a geometric interpretation of the derivative in terms of the slope of the secant lines of the graph of a real function  $f$  through  $[x_0, f(x_0)]$ . The derivative exists if and only if the slope of the secant line through the points  $[x_0, f(x_0)]$  and  $[x, f(x)]$  changes continuously when approaching the argument  $x = x_0$ . If so, the limit of this slope is the value of the derivative. This observation leads to the important corollary:

#### FUNCTIONS INCREASING AND DECREASING AT A POINT

**Corollary.** *If a real-valued function  $f$  has derivative  $f'(x_0) > 0$  at a point  $x_0 \in \mathbb{R}$ , then there is a neighbourhood  $\mathcal{O}(x_0)$  such that  $f(x) > f(x_0)$  for all points  $x \in \mathcal{O}(x_0)$ ,  $x > x_0$ , and  $f(x) < f(x_0)$  holds for all  $x \in \mathcal{O}(x_0)$ ,  $x < x_0$ .*

*On the other hand, if the derivative satisfies  $f'(x_0) < 0$ , then there is a neighbourhood  $\mathcal{O}(x_0)$  such that  $f(x) < f(x_0)$  for all points  $x \in \mathcal{O}(x_0)$ ,  $x > x_0$ , and  $f(x) > f(x_0)$  for all  $x \in \mathcal{O}(x_0)$ ,  $x < x_0$ .*

**PROOF.** Suppose  $f'(x_0) > 0$ . By the previous lemma,  $f(x) = f(x_0) + \psi(x)(x - x_0)$  and  $\psi(x_0) > 0$ . Since  $\psi$  is continuous at  $x_0$ , there exists a neighbourhood  $\mathcal{O}(x_0)$  on which  $\psi(x) > 0$ . If  $x$  increases,  $x > x_0$ ,  $f(x)$  increases as well,  $f(x) > f(x_0)$ . Analogously for  $x < x_0$ . The case with a negative derivative is proved similarly.  $\square$

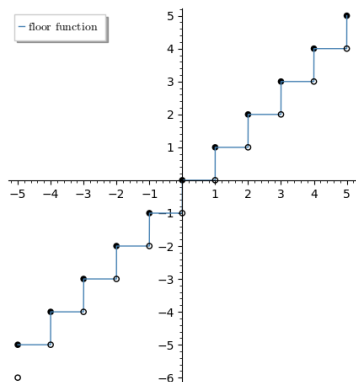
A real function is called *increasing at  $x_0$*  of its domain, if for all points  $x$  of some neighbourhood of a point  $x_0$ ,  $f(x) > f(x_0)$  if  $x > x_0$  and  $f(x) < f(x_0)$  if  $x < x_0$ . A real function is *increasing on an interval  $A$*  if  $f(x) - f(y) > 0$  for all  $x > y$ ,  $x, y \in A$ .

Similarly, a function is said to be *decreasing at a point  $x_0$*  if and only if there is a neighbourhood of the point  $x_0$  such that  $f(x) < f(x_0)$  for all  $x > x_0$ , while  $f(x) > f(x_0)$  for all  $x < x_0$  from this neighbourhood. A function is *decreasing on an interval  $A$*  if  $f(x) - f(y) < 0$  for all  $x > y$ ,  $x, y \in A$ .

Thus a function having a non-zero finite derivative at a point is either increasing or decreasing at that point, according to the sign of the derivative.

```
for x in [-5..5] :
    p+=point([x,x], size=30, color="black")
    p+=circle((x,x-1), 0.08, color="black")
show(p)
```

Sage prints out the figure given here (ignore the vertical lines, as they are not part of the graph):



Notice that there exist a shorter way to sketch the graph of the floor function, which combines the `floor` command with the command `exclude`. The latter automatically removes the vertical lines and the syntax goes as follows:

```
plot(floor(x), (x, -5, 5), exclude=[-5..5])
```

To solve problems involving the continuity of piecewise functions, there are generally two steps. First, prove that the individual components of the function are continuous within the intervals where the domain is divided. Second, examine the continuity at the “gluing points”. For the first step, you can often rely on the inherent continuity of elementary functions within their defined domains (e.g., polynomials, trigonometric functions, power functions, exponentials, logarithms, etc.).



**5.B.49.** Consider the 2-parameter family  $f_{\alpha,\beta}$  of piecewise functions, defined by

$$f_{\alpha,\beta}(x) = \begin{cases} \frac{\sqrt{2x^2 - x + 6} - \alpha x}{x + 2}, & \text{if } x < -2, \\ x^3 + \beta x + 4, & \text{if } x \geq -2, \end{cases}$$

with  $\alpha, \beta \in \mathbb{R}$ . Find the continuous members of  $f_{\alpha,\beta}$ . Next use Sage to confirm your answer, and moreover sketch the graph of these members, for  $-10 \leq x \leq 10$ .

The continuity of a function has far-reaching consequences, such as Bolzano’s theorem, which states that a continuous function which changes its sign within an interval has a zero, see 5.2.19. The following tasks are based on this important principle, with further examples provided in Section E.



A function increasing on an interval is increasing at each of its points. The converse is true as well. In order to see this, assume that  $f$  is increasing at all points of the interval  $A$ . Consider two points  $x < y$  in  $A$  with  $f(y) \leq f(x)$ . By the assumption, there is a  $\delta$ -neighbourhood of  $y$  on which  $z < y$  implies  $f(z) < f(y)$ . Let  $\delta_0$  be the supremum of all such  $\delta \leq y - x$  and  $w = y - \delta_0$ . Then  $f(w)$  cannot be larger than  $f(y)$  (there would be such a point on the right of it too, which is excluded). But, unless  $w = x$ ,  $w$  is a limit point of a sequence of points less than  $w$ , for which the value of  $f$  is larger than  $f(y) \geq f(w)$ . This is a contradiction with  $f$  increasing in  $w$ . But if  $w$  were  $x$  then there would be the contradiction with the assumption  $f(x) > f(z)$  for points  $z > x$  arbitrarily close to  $x$ .

The same arguments work for decreasing functions. The following is now proved:

**Proposition.** A real function is increasing or decreasing on an open interval  $A$  if and only if it is increasing or decreasing in each its point, respectively.

**5.3.3. Examples.** (1) There is a function which is increasing at the origin  $x_0 = 0$  but is neither increasing or decreasing on any neighbourhood of  $x_0$ . Consider the (continuous) function



$$f(x) = x + 5x^2 \sin(1/x), \quad f(0) = 0.$$

The choice  $f(0)$  makes  $f$  a continuous function on  $\mathbb{R}$  (sin is a bounded function with values between 1 and  $-1$ ). Its derivative at zero exists too.

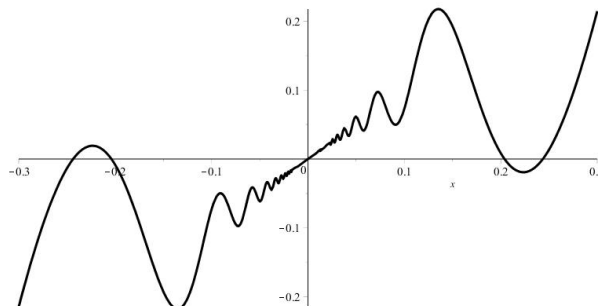
$$\lim_{x \rightarrow 0} \frac{x + 5x^2 \sin(1/x)}{x} = \lim_{x \rightarrow 0} (1 + 5x \sin(1/x)) = 1.$$

For  $x \neq 0$ ,

$$f'(x) = 1 + 10x \sin(1/x) - 5 \cos(1/x)$$

(cf. the rules for computing derivatives in 5.3.4 below). The derivative is not continuous at the origin.

$f$  is increasing at  $x = 0$  but is not increasing on any neighbourhood of this point.



(2) As another illustration of a simple usage of the relation between the derivatives and the properties of being an increasing (or decreasing) function, we can consider the existence of inverses to real polynomials.

Polynomials of degree at least two need not be either increasing or decreasing functions. Hence we cannot anticipate that there would be a globally defined inverse function to them. On the other hand, the inverse exists to every restriction

**5.B.50.** Determine whether the equation  $e^{2x} - x^4 + 3x^3 - 6x^2 = 5$  has a positive solution.

**Solution.** Let us consider the function

$$f(x) = e^{2x} - x^4 + 3x^3 - 6x^2 - 5, \quad x \geq 0,$$

for which  $f(0) = -4$ . Notice that  $\lim_{x \rightarrow +\infty} f(x) = \lim_{x \rightarrow +\infty} e^{2x} = +\infty$ . Obviously,  $f$  is continuous on the whole domain and hence it takes on all the values  $y \in [-4, \infty)$ . Moreover, we have  $f(0) < 0$  and  $f(2) = e^4 - 21 > 0$ . Hence by Bolzano's theorem its graph necessarily intersects the positive semi-axis  $x$ , i.e., the equation  $f(x) = 0$  has a positive solution. To confirm all these claims in Sage use the cell

```
f(x)=e^(2*x)-x^4+3*x^3-6*x^2-5
show(f(0)); show(f(2))
plot(f(x), x, 0, 2, ymax=20)
sol=solve(f(x)==0, x); show(sol)
```

which in fact solves the equation  $f(x) = 0$  via the last command. However, in this way Sage is not able to present the numerical expression of the complicated symbolic expression which returns for the solution (check it yourself). To avoid this problem, one may instead use the command

```
f(x)=e^(2*x)-x^4+3*x^3-6*x^2-5
find_root(f, 0, 2)
```

**5.B.51.** Determine whether the polynomial  $P(x) = x^{37} + 5x^{21} - 4x^9 + 5x^4 - 2x - 3$  has a real root in  $(-1, 1)$ . ○

**5.B.52.** Use Sage and implement Bolzano's theorem to prove that the equation

$$x^3 - \cos(x)e^x + x \sin(x) = 0$$

has a positive solution in  $(0, \pi/2)$ . Next use the `find_root` function to take a numerical approximation of this root. ○

### C. Derivatives

*Having both a conceptual understanding of the notion of limits but also the ability to compute limits, we are now ready to proceed with problems in calculus, the branch of mathematics devoted to derivatives and integrals of functions.<sup>6</sup> We will first focus on derivatives.*

*Given a function  $f$  depending on a real variable  $x$ , the slope of its graph at the point  $[x_0, f(x_0)]$  indicates the rate of change of the value  $f(x)$ , as  $x$  approaches the point  $x_0$ . This rate of change is the limiting value of the slopes  $\frac{f(x)-f(x_0)}{x-x_0}$  ( $x \neq x_0$ ), as  $x$  tends to  $x_0$ , that is,*

<sup>6</sup>A very rough notion of calculus, and in particular integration goes back to the ancient Greeks. Archimedes already developed the method of exhaustion to compute the volume of a cone, sphere, etc. However, calculus was essentially developed by Isaac Newton (1643–1727), and Gottfried Leibniz (1646–1716), both having as reference previous works by Fermat (1607–1665) and others. Later the French mathematician Augustin-Louis Cauchy (1789–1857) established a more rigorous approach to infinitesimal calculus, as we mentioned also before. On the other hand, the Italian mathematician Joseph-Louis Lagrange (1736–1813) was one of the founders of “calculus of variations”, whose elements will be studied in Chapter 9.

of  $f$  to an interval between adjacent roots of the derivative  $f'$ , i.e. where the derivative of the polynomial is non-zero and keeps the sign. These inverse functions will never be polynomials, except for the case of polynomials of degree one. The equation

$$y = ax + b$$

implies

$$x = \frac{1}{a}(y - b).$$

For a polynomial of degree two, the equation

$$y = ax^2 + bx + c$$

leads to the equation

$$x = \frac{-b \pm \sqrt{b^2 - 4a(c - y)}}{2a}.$$

Thus the inverse (given by the above equation) exists only for those  $x$  which are in either of the intervals  $(-\infty, -\frac{b}{2a})$ ,  $(-\frac{b}{2a}, \infty)$ .

It can be shown that the roots of polynomials of order larger than four cannot in general be expressed by means of power functions in general. Thus piece-wise defined inverses to polynomials may represent new items in our zoo.

### 5.3.4. Elementary properties of derivatives.

We introduce several basic facts about the calculation of derivatives. We shall see that the derivatives are quite nicely compatible with the algebraic operations of addition and multiplication of real or complex functions. The last formula then allows us to efficiently determine the derivatives of composite functions. It is also called the *chain rule*.

Intuitively, they can be understood very easily if we imagine that the derivative of a function  $y = f(x)$  is the quotient of the rates of increase of the output variable  $y$  and the input variable  $x$ :

$$f' = \frac{\delta y}{\delta x}.$$

Of course, for  $y = h(x) = f(x) + g(x)$ , the increase in  $y$  is given by the sum of the increases of  $f$  and  $g$ , and the increase of the input variable is still the same. Therefore, the derivative of a sum is the sum of the derivatives.

The derivative of a product is not the product of the derivatives. For  $y = f(x)g(x)$ , the increase is

$$\begin{aligned} \delta y &= f(x + \delta x)g(x + \delta x) - f(x)g(x) \\ &= f(x + \delta x)(g(x + \delta x) - g(x)) + (f(x + \delta x) - f(x))g(x) \end{aligned}$$

Now, if we make the increase  $\delta x$  small, we actually calculate the limit of a sum of products, which is the sum of the products of the limits. Thus the derivative of a product  $fg$  is given by the expression  $f'g + fg'$ , which is called the *Leibniz rule*<sup>10</sup>.

<sup>10</sup>Gottfried Wilhelm von Leibniz (1646-1716) was a great German mathematician and philosopher. He developed the differential and integral calculus in terms of the infinitesimal quantities, arguing similarly as above.



$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

If this limit exists and it is a finite number, then we say that  $f$  is “differentiable” at  $x_0$  and denote the limiting value by  $f'(x_0)$  or  $\frac{df}{dx}(x_0)$ .<sup>7</sup> This is the so-called first derivative of  $f$  at  $x_0$ .

The uniqueness of limits ensures that the derivative of  $f$  at  $x_0$ , when it exists, is uniquely determined. If  $f$  is differentiable at any point  $x_0 \in (a, b)$  in a given open interval, we say that  $f$  is differentiable on  $(a, b)$ . We say that  $f$  is differentiable on a closed interval  $[a, b]$ , when  $f$  is differentiable at any point  $x_0 \in (a, b)$  and the left-side and right-side derivatives at  $a$  and  $b$ , respectively, exist.

**5.C.1.** Based on the definition of the derivative of a function of one variable (see 5.3.1), compute the derivatives of the following functions:

$f(x) = x^n$ , with  $n \in \mathbb{N}$  constant and  $x \in \mathbb{R}$ ;

$g(x) = \sin(x)$ , with  $x \in \mathbb{R}$ ;

$h(x) = \sqrt{x}$ , with  $x \in [0, +\infty)$ . ○

**5.C.2.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a function such that  $f(a + b) = f(a)f(b)$ , for all  $a, b \in \mathbb{R}$ . Suppose also that  $f(0) = f'(0) = 1$ . Show that  $f'(x) = f(x)$  for all  $x \in \mathbb{R}$ . ○

**5.C.3.** Consider the function  $f(x) = |x|$ . Find the derivative of  $f$  for all  $0 \neq x \in \mathbb{R}$ . Is  $f$  differentiable at  $x_0 = 0$ ? ○

**5.C.4.** Show that if  $f$  is a differentiable function at a point  $x_0$  in its domain, then  $f$  is continuous at  $x_0$  as well. ○



The converse of the statement in 5.C.4 isn't true, and there are several reasons that a continuous function  $f$  may fail to be differentiable at a given point in its domain. A possible reason is a “corner” on its graph, as those of the absolute value function  $f(x) = |x|$  that we discussed in 5.C.3. Another reason is that  $f$  may admit a vertical tangent line at the given point, as the square root function discussed in 5.C.1

Let us summarize with a useful slogan: “Differentiability is a stronger requirement than continuity”.

**5.C.5.** Consider the function

$$f(x) = \begin{cases} \sqrt{x^2 + 7}, & \text{if } x > 3, \\ x^2 + bx + c, & \text{if } x \leq 3. \end{cases}$$

Determine the reals  $b, c \in \mathbb{R}$  such that  $f$  is differentiable for all  $x \in \mathbb{R}$ . For those values use Sage to plot the graph of  $f$  with at least two different methods.

**Solution.** Obviously, for all  $x \neq 3$  the given function is differentiable. In order to determine the unknowns  $b, c \in \mathbb{R}$  we need to establish a system of two equations. Since differentiability requires continuity, these equations are induced by the continuity and differentiability of  $f$  at  $x_0 = 3$ .

<sup>7</sup>For  $y = f(x)$ , the notation  $dy/dx$  is due to Leibniz, while  $f'(x)$  is the notation by Lagrange.

The derivative of a composite function is even more interesting: Consider a function

$$g = h \circ f,$$

where the domain of the function  $z = h(y)$  contains the codomain of the function  $y = f(x)$ . By writing out the increases,

$$g' = \frac{\delta z}{\delta x} = \frac{\delta z}{\delta y} \frac{\delta y}{\delta x}.$$

Thus we expect that the formula will be of the form

$$(h \circ f)'(x) = h'(f(x))f'(x).$$

Now we provide correct formulations together with proofs:

RULES FOR DIFFERENTIATION

**Theorem.** Let  $f$  and  $g$  be real or complex functions defined on a neighbourhood of a point  $x_0 \in \mathbb{R}$  and having finite derivatives at this point. Then

(1) for every real or complex number  $c$ , the function  $x \mapsto c \cdot f(x)$  has a derivative at  $x_0$  and

$$(cf)'(x_0) = c(f'(x_0)),$$

(2) the function  $f + g$  has a derivative at  $x_0$ , and

$$(f + g)'(x_0) = f'(x_0) + g'(x_0),$$

(3) the function  $f \cdot g$  has a derivative at  $x_0$ , and

$$(f \cdot g)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$

(4) Further, suppose  $h$  is a function defined on a neighbourhood of the image  $y_0 = f(x_0)$  with a derivative at  $y_0$ . Then the composite function  $h \circ f$  also has a derivative at  $x_0$ , and

$$(h \circ f)'(x_0) = h'(f(x_0)) \cdot f'(x_0).$$

**PROOF.** (1) and (2): A straightforward application of the theorem about sums and products of function limits yields the result.

(3) Rewrite the quotient of the increases (already mentioned), in the following way:

$$\frac{(fg)(x) - (fg)(x_0)}{x - x_0} = f(x) \frac{g(x) - g(x_0)}{x - x_0} + \frac{f(x) - f(x_0)}{x - x_0} g(x_0).$$

The limit of this as  $x \rightarrow x_0$  gives the desired result because  $f$  is continuous at  $x_0$ .

(4) By lemma 5.3.1, there are functions  $\psi$  and  $\varphi$  which are continuous at  $x_0$  and  $y_0 = f(x_0)$ . Further they satisfy  $h(y) = h(y_0) + \varphi(y)(y - y_0)$ ,  $f(x) = f(x_0) + \psi(x)(x - x_0)$  on some neighbourhoods of  $x_0$  and  $y_0$ . They satisfy  $\psi(x_0) = f'(x_0)$  and  $\varphi(y_0) = h'(y_0)$ . Then,

$$\begin{aligned} h(f(x)) - h(f(x_0)) &= \varphi(f(x))(f(x) - f(x_0)) \\ &= \varphi(f(x))\psi(x)(x - x_0) \end{aligned}$$

for all  $x$  from the neighbourhood of  $x_0$ . However, the product  $\varphi(f(x))\psi(x)$  is a function which is continuous at  $x_0$  and

We have  $f(3) = 9 + 3b + c$ , and the continuity condition  $\lim_{x \rightarrow 3^-} f(x) = 9 + 3b + c = \lim_{x \rightarrow 3^+} f(x)$ . The first equality holds as an identity and the second gives the relation  $9 + 3b + c = \sqrt{3^2 + 7} = 4$ , i.e.,  $3b + c = -5$ . As for the differentiability of  $f$  at  $x_0 = 3$  we get

$$\lim_{x \rightarrow 3^-} \frac{f(x) - f(3)}{x - 3} = \lim_{x \rightarrow 3^+} \frac{f(x) - f(3)}{x - 3}. \quad (*)$$

We see that  $f(3) = 9 + 3b + c = 9 - 5 = 4$  and hence

$$\begin{aligned} \lim_{x \rightarrow 3^-} \frac{f(x) - f(3)}{x - 3} &= \lim_{x \rightarrow 3^-} \frac{\sqrt{x^2 + 7} - 4}{x - 3} \\ &= \lim_{x \rightarrow 3^-} \frac{(\sqrt{x^2 + 7} - 4)(\sqrt{x^2 + 7} + 4)}{(x - 3)(\sqrt{x^2 + 7} + 4)} \\ &= \lim_{x \rightarrow 3^-} \frac{x^2 - 9}{(x - 3)(\sqrt{x^2 + 7} + 4)} \\ &= \lim_{x \rightarrow 3^-} \frac{x + 3}{\sqrt{x^2 + 7} + 4} = \frac{6}{8} = \frac{3}{4}. \end{aligned}$$

Similarly,

$$\begin{aligned} \lim_{x \rightarrow 3^+} \frac{f(x) - f(3)}{x - 3} &= \lim_{x \rightarrow 3^+} \frac{x^2 + b(x - 3) - 9}{x - 3} \\ &= \lim_{x \rightarrow 3^+} \frac{(x - 3)(x + 3 + b)}{x - 3} \\ &= \lim_{x \rightarrow 3^+} (x + 3 + b) = 6 + b. \end{aligned}$$

Thus, by (\*) we obtain the equation  $6 + b = \frac{3}{4}$ , that is,  $b = -\frac{21}{4}$  and so  $c = -5 - 3b = \frac{43}{4}$ .

In Sage we can plot  $f$  “manually”, i.e.,

```
f1(x)=sqrt(x^2+7); f2(x)=x^2-(21/4)*x+43/4
P=plot(f1(x), x, 3, 12, color="black")
P+=plot(f2(x), x, -6, 3, color="gray")
P+=point([3, 9-(21/4)*3+43/4], color="black",
size=30); show(P)
```

An alternative relies on the command `piecewise`, whose implementation can be summarized from the cell

```
f1(x)=sqrt(x^2+7); f2(x)=x^2-(21/4)*x+43/4
F=piecewise([[(-6, 3), f2(x)],
[(3, 12), f1(x)]])
p=plot(F(x), x, -6, 12)
p+=point([3, 9-(21/4)*3+43/4], color="black",
size=30); show(p)
```

Execute these cells to generate the required graph (note the code includes also the midpoint  $x_0 = 3$ ). Notably, compiling the “`piecewise`” method takes more time in both <https://sagecell.sagemath.org> or in [cocalc.com](http://cocalc.com), compared to the first approach. Finally, remember that another way to introduce  $f$  is based on the method presented in 5.B.49.  $\square$

its value at  $x_0$  is just the desired derivative of the composite function, again by lemma 5.3.1.  $\square$

**5.3.5. Derivative of quotients.** Consider first the special case of  $h(x) = x^{-1}$ . From the definition of the derivative,

$$\begin{aligned} h'(x) &= \lim_{\delta x \rightarrow 0} \frac{\frac{1}{x+\delta x} - \frac{1}{x}}{\delta x} = \lim_{\delta x \rightarrow 0} \frac{x - x - \delta x}{\delta x(x^2 + x\delta x)} \\ &= \lim_{\delta x \rightarrow 0} \frac{-1}{x^2 + x\delta x} = -x^{-2}. \end{aligned}$$

Thus, the above leads to:

DERIVATIVE OF A QUOTIENT

**Corollary.** Let  $f$  and  $g$  be real-valued functions which have finite derivatives at a point  $x_0$  and  $g(x_0) \neq 0$ . Then the function  $h(x) = f(x)(g(x))^{-1}$  satisfies

$$h'(x_0) = \left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2}.$$

**PROOF.** Using the formula  $(x^{-1})' = -x^{-2}$ , the chain rule says

$$(g^{-1})' = -g^{-2} \cdot g'.$$

The Leibniz rule implies

$$(f/g)' = (f \cdot g^{-1})' = f'g^{-1} - fg^{-2}g' = \frac{f'g - gf'}{g^2}.$$

$\square$

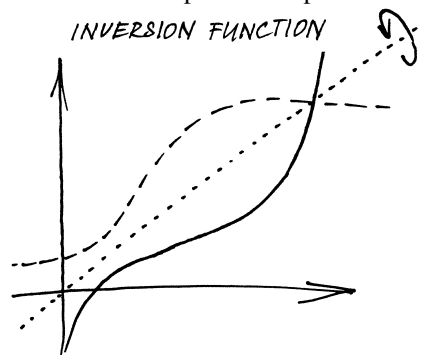
**5.3.6. Derivatives of inverse functions.** In paragraph 1.6.1,



while talking about relations and mappings in general, the concept of an *inverse function* was introduced. If the inverse function  $f^{-1}$  to a given function  $f : \mathbb{R} \rightarrow \mathbb{R}$  exists (do not confuse this notation with the function  $x \mapsto (f(x))^{-1}$ ), then it is uniquely determined by either of the following identities

$$f^{-1} \circ f = \text{id}_{\mathbb{R}}, \quad f \circ f^{-1} = \text{id}_{\mathbb{R}}.$$

Then the other identity is also true. If  $f$  is defined on a set  $A \subset \mathbb{R}$  and  $f(A) = B$ , the existence of  $f^{-1}$  is conditioned by the same statements with identity mappings  $\text{id}_A$  and  $\text{id}_B$ , respectively, on the right-hand sides. As seen in the diagram, the graph of the inverse function is obtained simply by interchanging the axes of the input and output variables.





The fundamental rules of differentiation (see 5.3.4 and 5.3.5) enable us to compute derivatives of functions, even those with complex forms. The upcoming exercises focus on these rules to enhance familiarity with their application. It's worth noting that while differentiation is a standard procedure, it can involve intricate computations. Therefore, many find it more enjoyable to compute derivatives using computer algebra packages, as these not only simplify the process but also allow for rigorous verification of calculations.

Below we will explain the usage of Sage. Similar syntaxes are available in many other computer algebra systems, as Mathematica and Maple.

**5.C.6.** In 5.C.1 we saw that  $(\sin(x))' \equiv \sin'(x) = \cos(x)$ . Based on this fact and using the *chain rule*, introduced in 5.3.4, prove that  $(\cos(x))' = -\sin(x)$ .  $\circ$

**5.C.7.** Differentiate the following functions, compute the required values and verify (some of) your answers via Sage.

- (1)  $f(x) = x \sin(x)$ . Find  $f'(\pi/2)$ ;
- (2)  $g(x) = \frac{\sin(x)}{x}$ ,  $x \neq 0$ . Find  $g'(\pi/2)$ ;
- (3)  $h(x) = \ln(x + \sqrt{x^2 - c^2})$ ,  $c \neq 0$ ,  $|x| \geq |c|$ .
- (4)  $k(x) = x^x$ ,  $x > 0$ ;
- (5)  $m(x) = x^{\sin(x)}$ ,  $x > 0$ .

**Solution.** (1) By the product rule (also called *Leibniz rule*, see 5.3.4), we see that

$$f'(x) = x' \cdot \sin(x) + x \cdot (\sin(x))' = \sin(x) + x \cos(x).$$

Thus  $f'(\pi/2) = 1$ . In order to differentiate a function  $f(x)$ , Sage provides many alternatives. For instance we can use the command `derivative(f,x)` or `f.derivative(x)`, as follows:

```
f(x)=x*sin(x); show(derivative(f, x))
```

or type

```
f(x)=x*sin(x); show(f.derivative(x))
```

If we want to find the explicit value of the derivative at a point  $x_0$ , then the command `f.derivative(x)(x=x_0)` is an appropriate one. Thus, to compute  $f'(\pi/2)$  you may continue typing in the previous cell the following:

```
f.derivative(x)(x=pi/2)
```

This prints out 1, that is,  $f'(\pi/2) = 1$ , where  $f(x) = x \sin(x)$ . An alternative to differentiate  $f(x)$  relies on the command `diff`. For example, by typing

```
f(x)=x*sin(x); show(diff(f, x))
```

we obtain the same result. In order to specify the value at a certain point  $x_0$ , we again type `diff(f,x)(x=x_0)`, i.e.,

```
f(x)=x*sin(x); show(diff(f, x)(x=pi/2))
```

The command `diff` is also useful when one wants to find higher-order derivatives, hence we will meet it again in Chapter 6, see also below and the final Section E of this chapter.

If it is known that the inverse  $y = f^{-1}(x)$  of a differentiable function  $x = f(y)$  is also differentiable, then the chain rule yields immediately

$$1 = (\text{id})'(x) = (f \circ f^{-1})'(x) = f'(y) \cdot (f^{-1})'(x).$$

Notice that  $f'(y)$  must be non-zero.

This corresponds to the intuitive idea that for  $y = f(x)$ , the value of  $f'$  is approximately  $\frac{\delta y}{\delta x}$  while for  $x = f^{-1}(y)$  it is approximately  $(f^{-1})'(y) = \frac{\delta x}{\delta y}$ . And this indeed is the way the derivatives of inverse functions are calculated.

#### DERIVATIVE OF THE INVERSE FUNCTION

**Theorem.** If  $f$  is a real-valued function differentiable at  $y_0$ , such that the inverse  $f^{-1}(x)$  exists on a neighbourhood of the value  $x_0 = f(y_0)$  and  $f'(y_0) \neq 0$ , then

$$(1) \quad (f^{-1})'(x_0) = \frac{1}{f'(f^{-1}(x_0))} = \frac{1}{f'(y_0)}.$$

Let us notice that if we assume that the derivative  $f'(x)$  is continuous on a neighborhood of  $x_0$ , then the condition  $f'(x_0) \neq 0$  clearly implies that  $f$  is either increasing or decreasing on some neighborhood, and thus the assumptions of the theorem are fulfilled. On the other hand, the example 5.3.3(1) shows, that the existence of  $f'(x_0) \neq 0$  is not necessarily sufficient.

**PROOF.** To prove the proposition, it suffices to read the proof of the fourth statement of theorem 5.3.4. We work with the composition  $f \circ f^{-1} = \text{id}$  there. The composite function is differentiable. By lemma 5.3.1, there is a function  $\varphi$  continuous at  $y_0$  such that



$$f(y) - f(y_0) = \varphi(y)(y - y_0),$$

on some neighbourhood of  $y_0$ . Further, it satisfies  $\varphi(y_0) = f'(y_0) \neq 0$  and  $\varphi$  has constant sign on a neighbourhood of  $x_0$ . Next, notice that the existence of the inverse  $f^{-1}$  around the point  $x_0$  and the continuity of  $f$  at  $y_0$  guarantees the continuity of  $f^{-1}$  at  $x_0$ , (the  $\varepsilon$  and  $\delta$  neighbourhoods map each to the other bijectively). The substitution  $y = f^{-1}(x)$  then yields

$$x - x_0 = \varphi(f^{-1}(x))(f^{-1}(x) - f^{-1}(x_0)),$$

for all  $x$  lying in some neighbourhood  $\mathcal{O}(x_0)$  of  $x_0$ . Further,  $f^{-1}(x_0) = y_0$ , and  $\varphi(f^{-1}(x))$  is continuous at  $x_0$  and remains non-zero on a neighbourhood  $\mathcal{O}(x_0)$  of  $x_0$  with constant sign. Thus

$$\frac{f^{-1}(x) - f^{-1}(x_0)}{x - x_0} = \frac{1}{\varphi(f^{-1}(x))} \neq 0,$$

for all  $x \in \mathcal{O}(x_0) \setminus \{x_0\}$ . The right-hand side of this expression is continuous at  $x_0$ . The limit is

$$\lim_{x \rightarrow x_0} \frac{1}{\varphi(f^{-1}(x))} = \frac{1}{\varphi(f^{-1}(x_0))} = \frac{1}{f'(y_0)}.$$

(2) The rule for the derivative of a quotient (see 5.3.5), gives

$$g'(x) = \frac{(\sin(x))' \cdot x - \sin(x) \cdot x'}{x^2} = \frac{x \cos(x) - \sin(x)}{x^2}.$$

Thus, we easily get  $g'(\pi/2) = -\frac{4}{\pi^2}$ . To confirm this via Sage a recommendation has the form

```
g(x)=sin(x)/x; diff(g(x), x)(x=pi/2)
```

(3) Here the “chain rule” applies. Set  $a(x) = \ln(x)$  and  $b(x) = x + \sqrt{x^2 - c^2}$  such that  $h(x) = a(b(x))$ . Then

$$\begin{aligned} h'(x) &= (a(b(x)))' = a'(b(x)) \cdot b'(x) \\ &= \frac{(x + \sqrt{x^2 - c^2})'}{x + \sqrt{x^2 - c^2}} = \frac{1 + \frac{x}{x^2 - c^2}}{x + \sqrt{x^2 - c^2}}. \end{aligned}$$

For a confirmation in Sage you may use one of the methods posed above.

(4) Let us recall the identity  $(e^{f(x)})' = f'(x)e^{f(x)}$ , where we assume that  $f'(x)$  exists. Because  $x^x = e^{x \ln(x)}$  by applying this rule we obtain

$$\begin{aligned} k'(x) = (x^x)' &= (e^{x \ln(x)})' = (x \ln(x))' e^{x \ln(x)} \\ &= (\ln(x) + 1) e^{x \ln(x)} = (\ln(x) + 1)x^x. \end{aligned}$$

In Sage you may use the cell `show(diff(x**x, x))`, which returns the expression  $x^x(\log(x) + 1)$  (recall that in Sage the function `log` represents the natural logarithm).

The same trick used in (4) applies for the final case, too. Hence we leave this as an exercise.  $\square$

**5.C.8. Chain rule via Sage.** Write in Sage a short code implementing the chain rule for two arbitrary differentiable functions.

**Solution.** Let  $h(x) = f(g(x))$  be the composition of two arbitrary differentiable functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$ . We want to implement in Sage the rule  $h'(x) = g'(x)f'(g(x))$ , for all  $x \in \mathbb{R}$ , and with this goal in mind it is useful to recall about symbolic functions in Sage. For instance, type in your editor the cell

```
x=var("x")
f=function('f')(x); show(f)
```

This cell introduces an arbitrary function  $f$  of one variable  $x$ , in particular prints out  $f(x)$ . This makes the implementation of the chain rule for two arbitrary functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$ , really simple. First introduce the composition  $h$ , as follows

```
x=var("x"); f=function("f"); g=function("g")
h=f(g(x)) # define the composition h
show(h)
```

Check yourself that this gives the composition  $f(g(x))$ . Next it is sufficient to add the command `h.diff(x)`, which returns  $D[0](f)(g(x)) * diff(g(x), x)$ . Recall here that  $D[0]$  encodes the first derivative, hence we can indeed translate Sage’s output as  $f'(g(x))g'(x)$ , that is, the chain rule.

A more useful alternative has the following form:

```
function("f")(x)
```

Therefore, the limit of the left-hand side also exists, and it follows that

$$(f^{-1})'(x_0) = \frac{1}{f'(y_0)}$$

as required.  $\square$

**5.3.7. Derivatives of the elementary functions.** Consider the exponential function  $f(x) = a^x$  for any fixed real  $a > 0$ . If the derivative of  $a^x$  exists for all  $x$ , then

$$f'(x) = \lim_{\delta x \rightarrow 0} \frac{a^{x+\delta x} - a^x}{\delta x} = a^x \lim_{\delta x \rightarrow 0} \frac{a^{\delta x} - 1}{\delta x} = f'(0)a^x.$$

On the other hand, if the derivative at zero exists, then this formula guarantees the existence of the derivative at any point of the domain and also determines its value. At the same time, the validity of this formula for one-sided derivatives is also verified.

Unfortunately, it takes some time to verify that the derivatives of exponential functions indeed exist (see 5.4.2, 5.4.10, and 6.3.7).

There is an especially important base  $e$ , also known as Euler’s number, for which the derivative at zero equals one.

Remember the formula  $(e^x)' = e^x$  for a while and draw on its consequences.

For the general exponential function, (using standard rules of differentiation),

$$(a^x)' = (e^{\ln(a)x})' = \ln(a)(e^{\ln(a)x}) = \ln(a) \cdot a^x.$$

Thus exponential functions are special since their derivatives are proportional to their values.

Next, we determine the derivative  $(\ln_e(x))'$ . The definition of the natural logarithm as the inverse to  $e^x$ ,

$$e^{\ln x} = x,$$

allows the calculation:

$$(1) \quad (\ln)'(y) = (\ln)'(e^x) = \frac{1}{(e^x)'} = \frac{1}{e^x} = \frac{1}{y}.$$

The formula

$$(2) \quad (x^a)' = ax^{a-1}$$

for differentiating a general power function can also be derived using the derivatives of the exponential and logarithmic functions:

$$(x^a)' = (e^{a \ln x})' = e^{a \ln x} (a \ln x)' = a \frac{x^a}{x} = ax^{a-1}.$$

**5.3.8. Mean value theorems.** Before continuing the journey of finding new interesting functions, we derive several simple statements about derivatives. The meaning of all of them is intuitively clear from the diagrams. The proofs follow the visual imagination.





```
function("g")(x)
show(diff(f(g(x)), x))
```

The output here has the form  $D_0(f)(g(x)) \frac{\partial}{\partial x} g(x)$ , and the advantage is that we can “replace”  $f, g$  with certain functions, so that we can use this method for verifying our formal computations in such examples. For instance, to compute the derivative of  $e^{x^2}$ , it is sufficient to specify  $f, g$  by adding in the previous cell the code

```
f(x)=e^x;g(x)=x^2; show(diff(f(g(x)), x))
```

To compute the case (5) of Problem 5.C.7, add the line  $f(x) = x^x; g(x) = \sin(x); show(diff(f(g(x)), x))$ . Try some additional examples yourself, and see also in Chapter 6 for further applications of symbolic functions (cf. 6.A.4).  $\square$

**5.C.9. Inverse function theorem.** (a) Given the function  $f(x) = (4x + 3)/(x - 6)$  prove that it is invertible on its domain, with inverse given by the function  $g(x) = (6x + 3)/(x - 4)$ , with  $x \in \mathbb{R} \setminus \{4\}$ .

(b) According to 5.3.6 if  $y = g(x)$  is the inverse of a differentiable function  $f(x)$ , then for all  $x$  satisfying  $f'(g(x)) \neq 0$  we have  $g'(x) = 1/f'(g(x))$ . Based on this result compute the derivative of the function  $g(x)$  given in (a) and then compare your result, by applying a direct computation of  $g'(x)$ .

**Solution.** (a) Obviously, the domain of  $f$  is the set  $A := \mathbb{R} \setminus \{6\}$ . For arbitrary  $x_1, x_2 \in A$  we see that the relation  $f(x_1) = f(x_2)$  is equivalent to  $27x_1 = 27x_2$ , that is,  $x_1 = x_2$ . Hence  $f$  is one-to-one and its inverse  $y = f^{-1}(x)$  exists. We will show that  $y = f^{-1}(x) = g(x)$  with domain the set  $B = \mathbb{R} \setminus \{4\}$ . Setting  $y = \frac{4x+3}{x-6}$ , we equivalently get  $y(x-6) = 4x+3$ , or  $x(y-4) = 6y+3$ , that is  $x = \frac{6y+3}{y-4}$  for  $y \neq 4$ . Such  $x$  are indeed elements of  $A$ , and in order to obtain the inverse of  $f$  it is now sufficient to reverse the roles of  $x, y$  in the previous relation. This gives the desired expression of  $g = f^{-1}$ . An alternative verification occurs due to the relations  $f(g(x)) = \frac{27x}{27} = x$  and  $g(f(x)) = \frac{27x}{27} = x$ . (b) A direct computation shows that

$$g'(x) = \left(\frac{6x+3}{x-4}\right)' = -\frac{27}{(x-4)^2}, \quad x \in B.$$

Let us obtain the same result by applying the method mentioned above. We similarly compute  $f'(x) = -\frac{27}{(x-6)^2}$  and hence it follows that for all  $x \neq 4$  we have

$$f'(g(x)) = -\frac{27}{\left(\frac{6x+3}{x-4} - 6\right)^2} = -\frac{(x-4)^2}{27} \neq 0.$$

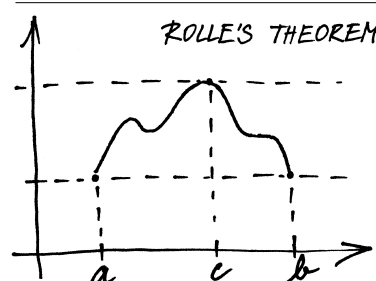
Thus  $g'(x) = 1/f'(g(x)) = -\frac{27}{(x-4)^2}$ , for all  $x \in B$ .  $\square$

Let  $y = f(x)$  be a differentiable function at  $x_0 \in \mathbb{R}$ . The tangent line of the graph of  $f$  at the point  $[x_0, y_0 = f(x_0)] \in \mathbb{R}^2$  has the form  $y - y_0 = f'(x_0)(x - x_0)$ , or equivalently

$$y = f'(x_0)x + (y_0 - f'(x_0)x_0).$$

### ROLLE'S THEOREM<sup>11</sup>

**Theorem.** Assume that the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is continuous on a closed bounded interval  $[a, b]$  and differentiable inside this interval. If  $f(a) = f(b)$ , then there is a number  $c \in (a, b)$  such that  $f'(c) = 0$ .



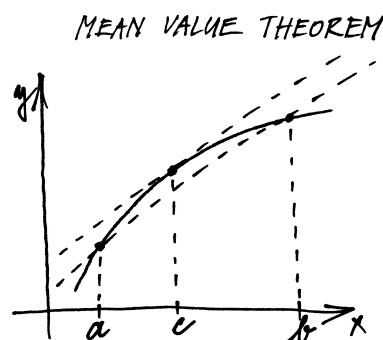
**PROOF.** Since the function  $f$  is continuous on the closed interval (i.e. on a compact set), it attains its maximum and its minimum there. Either its maximum value is greater than  $f(a) = f(b)$ , or the minimum value is less than  $f(a) = f(b)$ , or  $f$  is constant. If the third case applies, the derivative is zero at all points of the interval  $(a, b)$ . If the second case applies, then the first case applies to the function  $-f$ . If the first case applies, it occurs at an interior point  $c$ . If  $f'(c) \neq 0$  then the function  $f$  would be either increasing or decreasing at  $c$  (see 5.3.2), implying the existence of larger values than  $f(c)$  in a neighbourhood of  $c$ , contradicting that  $f(c)$  is a maximum value.  $\square$

**5.3.9.** The latter result immediately implies the following corollary.

### LAGRANGE'S MEAN VALUE THEOREM

**Theorem.** Assume the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is continuous on an interval  $[a, b]$  and differentiable at all points inside this interval. Then there is a number  $c \in (a, b)$  such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$



<sup>11</sup>The French mathematician Michel Rolle (1652-1719) proved this theorem only for polynomials. The principle was perhaps known much earlier, but the rigorous proof comes from the 19th century only.

Hence, the tangent is the line through the point  $[x_0, y_0]$ , having slope  $f'(x_0)$ . Observe that the geometric condition of having a unique non-vertical tangent to the graph of  $f$  at a point  $[x_0, f(x_0)] \in G_f$  is equivalent to the existence of  $f'(x_0)$ , that is, to the differentiability of  $f$  at  $x_0$ . Hence tangent lines provide a more intuitive definition of differentiability: We may always think a differentiable function  $f$  at  $x_0$ , as a function whose graph has a unique (non-vertical) tangent at the point  $[x_0, f(x_0)]$ . In this case the graph of  $f$  cannot have breaks, corners or cusps at  $x_0$ .



PROOF. The proof is a simple statement of the geometrical meaning of the theorem: The secant line between the points  $[a, f(a)]$  and  $[b, f(b)]$  has a tangent line which is parallel to it (have a look at the diagram). The equation of the secant line is

$$y = g(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a).$$

The difference  $h(x) = f(x) - g(x)$  determines the (vertical) distance of the graph and the secant line (in the values of  $y$ ). Surely  $h(a) = h(b) = 0$  and

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

By Rolle's theorem, there is a point  $c$  at which  $h'(c) = 0$ .  $\square$

**5.C.10.** Consider the function  $f(x) = \alpha x^2 - 4x \ln(x)$  with  $x > 0$ , where  $\alpha \neq 0$  is some real parameter. Find the tangent of  $f$  at the point  $P = [1, f(1)]$ . Next specify  $\alpha$  such that the origin is a point of this tangent.  $\square$

**5.C.11.** Consider the function  $f(x) = x^2 \ln(x)$ , with  $x > 0$ . Prove that there exists a unique point  $P \in \mathbb{R}^2$  belonging on the graph of  $f$  and where the tangent of  $f$  is parallel to the  $x$ -axis. Explain why this happens and verify your computations via Sage.

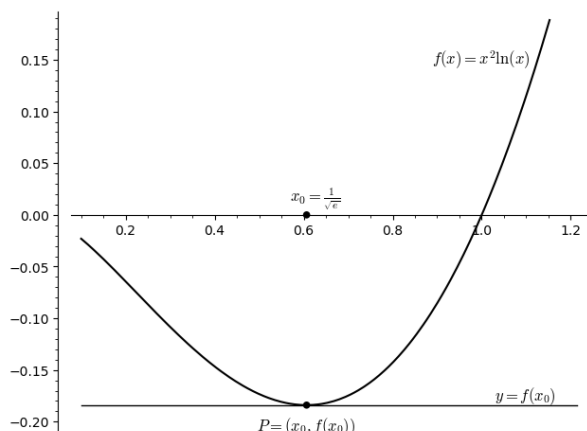
**Solution.** The derivative of  $f$  is given by

$$f'(x) = 2x \ln(x) + x = x(2 \ln(x) + 1),$$

for any  $x \in (0, +\infty)$ . We see that the equation  $f'(x) = 0$  has a unique solution, namely  $x_0 = e^{-\frac{1}{2}} = \frac{1}{\sqrt{e}}$  (since  $f'$  is defined only for  $x > 0$ , the solution  $x = 0$  is not acceptable). In Sage we can verify these computations by typing

```
f(x)=x^2*ln(x); assume(x>0)
show(solve(diff(f, x)==0, x))
```

The tangent line of  $f$  at this point has zero slope, i.e.,  $f'(x_0) = 0$ , and hence is horizontal. In particular, the tangent line is given by  $y = f(x_0)$ , and observe that near  $x_0$  this line lies under the graph of  $f$ . In such a situation we say that at  $x_0$  the function  $f$  attains a local minimum with value  $f(x_0)$ , see also here:



Thus there is a unique point  $P \in \mathbb{R}^2$  satisfying the statement, with coordinates  $P = \left[ \frac{1}{\sqrt{e}}, -\frac{1}{2e} \right]$ .  $\square$

The mean value theorem can also be written in the form:

$$(1) \quad f(b) = f(a) + f'(c)(b - a).$$

In the case of a parametrically given curve in the plane, i.e., a pair of functions  $y = f(t)$ ,  $x = g(t)$ , similar result about the existence of a tangent line parallel to the secant line going through the boundary points is described by *Cauchy's mean value theorem*. Notice we may consider such a curve as a complex valued function  $f(t) + i g(t)$ .

#### CAUCHY'S MEAN VALUE THEOREM

**Corollary.** Let functions  $y = f(t)$  and  $x = g(t)$  be continuous on an interval  $[a, b]$  and differentiable inside this interval. Further, let  $g'(t) \neq 0$  for all  $t \in (a, b)$  and  $g(b) \neq g(a)$ . Then there is a point  $c \in (a, b)$  such that

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)}.$$

PROOF. Put

$$h(t) = (f(b) - f(a))g(t) - (g(b) - g(a))f(t).$$

Now  $h(a) = f(b)g(a) - f(a)g(b)$ ,  $h(b) = f(b)g(b) - f(a)g(b)$ , so by Rolle's theorem, there is a number  $c \in (a, b)$  such that  $h'(c) = 0$ .

Finally,  $g'(c) \neq 0$  and the desired formula follows.  $\square$

Notice that  $g(b) \neq g(a)$  automatically, if  $g'(t)$  is continuous.

**5.3.10.** A reasoning similar to the one in the above proof leads to a supremely useful tool for calculating limits of quotients of functions.

**5.C.12.** Find the equations of the tangent and normal lines to the graph of  $g(x) = (x + 1) \sqrt[3]{3 - x}$ , with  $x \in \mathbb{R}$ , at the point  $P = [1, f(1)]$ . Next sketch the two lines together with the graph of  $g$  via Sage.

**Solution.** We have  $g(x_0) = g(1) = 2 \cdot 2^{\frac{1}{3}} = 2^{\frac{4}{3}}$ . Moreover

$$g'(x) = (3 - x)^{\frac{1}{3}} - \frac{1}{3}(x + 1)(3 - x)^{-\frac{2}{3}},$$

and hence  $g'(1) = \frac{2}{3} \cdot 2^{\frac{1}{3}} = \frac{2\sqrt[3]{2}}{3}$ . Thus the tangent of  $g$  at  $P$  is given by

$$y_1 = g(x_0) + g'(x_0)(x - x_0) = 2^{\frac{4}{3}} + \frac{2\sqrt[3]{2}}{3}(x - 1).$$

The normal line and tangent line are perpendicular, hence they should have slopes that are opposite reciprocals each other. This means that the normal line has the form

$$y_2 = g(x_0) - \frac{1}{g'(x_0)}(x - x_0) = 2^{\frac{4}{3}} - \frac{3}{2\sqrt[3]{2}}(x - 1).$$

In order to derive the equations of the tangent and normal (and sketch them) in Sage, type the block

```
g(x)=(x+1)*(3-x)^(1/3);dg(x)=diff(g(x),x)
a=plot(g(x),x,-1.2,3,xmin=-3,xmax=3,
legend_label="curve")
x0=1;tangl=g(x0)+dg(x0)*(x-x0)
show(tangl)
perpl=g(x0)-(1/dg(x0))*(x-x0);show(perpl)
b=plot(tangl,xmin=-1.2,xmax=1.5,
color="black",
legend_label="tangent",aspect_ratio=1)
c=plot(perpl,xmin=-1,xmax=2,color="gray",
legend_label="normal")
m=point([x0,0],size=30,color="black")
mv=point([x0,g(x0)],size=30,color="black")
show(a+b+c+m+mv)
```

We leave to the reader the implementation of this block, for practicing with Sage.  $\square$

**5.C.13.** Find the tangent and normal line to the curve given by the equation  $x^3 + y^3 - 2xy = 0$ , at the point  $P = [1, 1]$ .  $\circ$

**5.C.14. Tangent lines via Sage.** Given the polynomial  $P(x) = x^4 - 2x$  with  $x \in I = [-2, 2]$ , write a short routine in Sage constructing the tangent line  $\ell_{x_0}(x)$  of  $P$  at a random point  $x_0$  lying in the interval  $I$ . Then choosing certain  $x_0 \in I$ , produce the graphs of  $P$  and  $\ell_{x_0}$  for all  $x \in I$ .

**Solution.** Using the `def` command we can introduce a sub-routine, which we agree to call `Tangent`. To do so we first need to introduce  $P$  and its first derivative. The whole block has the form

L'HOSPITAL'S RULE<sup>12</sup>

**Theorem.** Suppose  $f$  and  $g$  are functions differentiable on some neighbourhood of a point  $x_0 \in \mathbb{R}$ , yet not necessarily at  $x_0$  itself. Suppose

$$\lim_{x \rightarrow x_0} f(x) = 0, \quad \lim_{x \rightarrow x_0} g(x) = 0.$$

If the limit

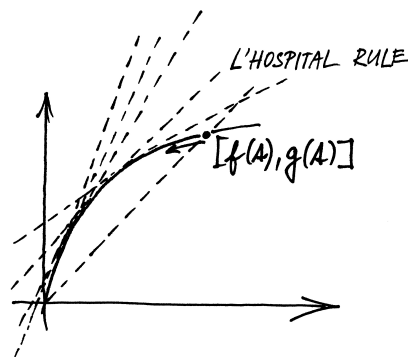
$$\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$$

exists, then the limit

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)}$$

also exists, and the two limits are equal.

**PROOF.** Without loss of generality, the functions  $f$  and  $g$  are zero at the point  $x_0$ . The quotient of the values then corresponds to the slope of the secant line between the points  $[0, 0]$  and  $[f(x), g(x)]$ . At the same time, the quotient of the derivatives corresponds to the slope of the tangent line at the given point. Thus it is necessary to verify that the limit of the slopes of the secant lines exists from the fact that the limit of the slopes of the tangent lines exists.



Technically, we can use the mean value theorem in Cauchy's parametric form. First of all, the existence of the expression  $f'(x)/g'(x)$  on some neighbourhood of the point  $x_0$  (excluding  $x_0$  itself) is implicitly assumed. Thus especially for points  $c$  sufficiently close to  $x_0$ ,  $g'(c) \neq 0$ .<sup>13</sup> By the mean value theorem,

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{g(x) - g(x_0)} = \lim_{x \rightarrow x_0} \frac{f'(c_x)}{g'(c_x)},$$

<sup>12</sup>Guillaume François Antoine, Marquis de l'Hôpital, (1661-1704) became famous for his textbook on Calculus (in modern textbooks, his name is mostly spelled as l'Hospital). This rule was first published there, perhaps originally proved by one of the famous Bernoulli brothers.

<sup>13</sup>This is not always necessary for the existence of the limit in a general sense. Nevertheless, for the statement of l'Hospital's rule, it is. A thorough discussion can be found (googled) in the popular article 'R. P. Boas, Counterexamples to L'Hospital's Rule, The American Mathematical Monthly, October 1986, Volume 93, Number 8, pp. 644-645.'

```
P(x)=x^4-2*x; P1(x)=diff(P(x), x)
def Tangent(x_0):
    y_0=P(x_0)
    m=P1(x_0)
    c=y_0 - m*x_0
    l(x)=m*x+c #this defines the tangent line
    Q = plot(P(x), -2, 2, color="blue",
```

where  $c_x$  is a number lying between  $x_0$  and  $x$ , dependent on  $x$ . From the existence of the limit

$$\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)},$$

it follows that this value will be shared by the limit of any sequence created by substituting the values  $x = x_n$  approaching  $x_0$  into  $f'(x)/g'(x)$  (cf. the convergence test 5.2.15). Especially, we can substitute any sequence  $c_{x_n}$  for  $x_n \rightarrow x_0$ , and thus the limit

$$\lim_{x \rightarrow x_0} \frac{f'(c_x)}{g'(c_x)}$$

exist, and the last two limits are equal. Hence the desired limit exists and has the same value.  $\square$

From the proof of the theorem, it is true for one-sided limits as well.

**5.3.11. Corollaries.** L'Hospital's rule can easily be extended for limits at the improper points  $\pm\infty$  and for the case of infinite values of the limits. If, for instance, we have

$$\lim_{x \rightarrow \infty} f(x) = 0, \quad \lim_{x \rightarrow \infty} g(x) = 0,$$

then  $\lim_{x \rightarrow 0^+} f(1/x) = 0$  and  $\lim_{x \rightarrow 0^+} g(1/x) = 0$ .

At the same time, from existence of the limit of the quotient of the derivatives at infinity,

$$\begin{aligned} \lim_{x \rightarrow 0^+} \frac{(f(1/x))'}{(g(1/x))'} &= \lim_{x \rightarrow 0^+} \frac{f'(1/x)(-1/x^2)}{g'(1/x)(-1/x^2)} \\ &= \lim_{x \rightarrow 0^+} \frac{f'(1/x)}{g'(1/x)} = \lim_{x \rightarrow \infty} \frac{f'(x)}{g'(x)}. \end{aligned}$$

Applying the previous theorem, the limit

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = \lim_{x \rightarrow 0^+} \frac{f(1/x)}{g(1/x)} = \lim_{x \rightarrow \infty} \frac{f'(x)}{g'(x)}$$

exists in this case as well.

The limit calculation is even simpler in the case when

$$\lim_{x \rightarrow x_0} f(x) = \pm\infty, \quad \lim_{x \rightarrow x_0} g(x) = \pm\infty.$$

Then it suffices to write

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow x_0} \frac{1/g(x)}{1/f(x)},$$

which is already the case of usage of l'Hospital's rule from the previous theorem.

In fact, the l'Hospital's rule has the same form for improper limits as well:

**Theorem.** Let  $f$  and  $g$  be functions differentiable on some neighbourhood of a point  $x_0 \in \mathbb{R}$ , not necessarily at  $x_0$  itself. Further, let the limits  $\lim_{x \rightarrow x_0} f(x) = \pm\infty$  and  $\lim_{x \rightarrow x_0} g(x) = \pm\infty$  exist. If the limit

$$\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)}$$

exists, then the limit

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)}$$

also exists and they equal each other.

```

ymin=-2, ymax=20, gridlines="true")
Q+=plot(l(x), -2, 2, color="black",
ymin=-2, ymax=20)
Q+=point((x_0, y_0), color="black",
size=50)
Q.show()
show("x_0=" , x_0)
show("tang. line: l(x)=(", m, ") *x+(", c, ")")
return

```

Notice that  $x_0$  is included in this code as a real number input. Also, the command `gridlines` with the option `true` was used to add grid lines in the background (though not necessary). The last two `show` commands are included to display the chosen value of  $x_0$  and the expression of the tangent line  $\ell_{x_0}(x)$  at  $x_0$ . Finally, the command `return` is used to conclude the routine. Further applications of programming in Sage will be discussed later, as seen, for for example in 5.D.5.

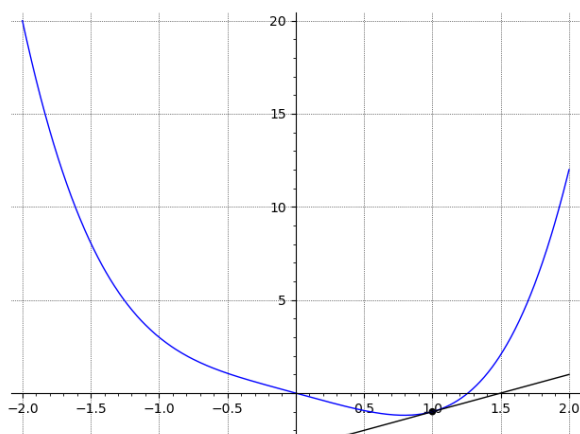
We can now check our routine by testing certain values for  $x_0$ . We may type

```

Tangent(0); Tangent(-0.5)
Tangent(0.5); Tangent(1)

```

etc. For instance, the last command produces the figure



and prints

```

x_0=1
tang. line: l(x)=(2)*x+(-3)

```

Later in the final section of this chapter we will use the routine constructed here to build an interactive environment for tangent lines, see 5.E.92.  $\square$

*We have now only scratched the surface of the theory of derivatives, which encompasses a wide range of theorems and applications. Derivatives offer a powerful mathematical framework that naturally applies to the study of monotone functions, optimization problems, and approximation techniques.*



**PROOF.** Apply the mean value theorem. The key step is to express the quotient in a form where the derivative arises:

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(y)}{f(x) - f(y)} \cdot \frac{f(x) - f(y)}{g(x) - g(y)} \cdot \frac{g(x) - g(y)}{g(x)},$$

where  $y \neq 0$  is fixed, from a neighbourhood of  $x_0$  and  $x$  is approaching  $x_0$ . Since the limits of  $f$  and  $g$  at  $x_0$  are infinite, we can surely assume that the differences of the values of both functions at  $x$  and  $y$ , having fixed  $y$ , are non-zero.

Using the mean value theorem, replace the fraction in the middle with the quotient of the derivatives at an appropriate point  $c$  between  $x$  and  $y$ . The expression of the examined limit thus gets the form



$$\frac{f(x)}{g(x)} = \frac{1 - \frac{g(y)}{g(x)}}{1 - \frac{f(y)}{f(x)}} \cdot \frac{f'(c)}{g'(c)},$$

where  $c$  depends on both  $x$  and  $y$ . With  $y$  fixed and  $x$  approaching  $x_0$ , the former fraction converges to one. At the same time, if simultaneously  $y \rightarrow x_0$  and  $|y - y_0| \geq |x - x_0|$ , the latter fraction becomes arbitrarily close to the limit value of the quotient of the derivatives. Thus, we may choose a sequence  $y_n \rightarrow x_0$ , such that  $|\lim_{x \rightarrow x_0} \frac{f'(x)}{g'(x)} - \frac{f'(c)}{g'(c)}| < 1/n$  for all  $c$  with  $|c - x_0| < |y_n - x_0|$ . At the same time, we may restrict appropriately  $|x - x_0| \leq \delta_n$ , so that the former fraction with  $y = y_n$  will be closer to 1 than by  $1/n$  for all such  $x$ . Altogether, this implies the requested equality of the limits.  $\square$

By making suitable modifications of the examined expressions, one can also apply l'Hopital's rule on forms of the types  $\infty - \infty$ ,  $1^\infty$ ,  $0 \cdot \infty$ , and so on. Often one simply rearranges the expressions or uses some continuous function, for instance the exponential one.



**5.3.12. Example.** For an illustration of such a procedure, we show the connection between the arithmetic and geometric means of  $n$  non-negative values  $x_i$ . The *arithmetic mean*

$$M^1(x_1, \dots, x_n) = \frac{x_1 + \dots + x_n}{n}$$

is a special case of the *power mean with exponent  $r$* , also known as the generalized mean:

$$M^r(x_1, \dots, x_n) = \left( \frac{x_1^r + \dots + x_n^r}{n} \right)^{\frac{1}{r}}.$$

The special value  $M^{-1}$  is called the *harmonic mean*. Calculate the limit value of  $M^r$  for  $r$  approaching zero. For this purpose, determine the limit by l'Hopital's rule (we treat it as an expression of the form  $0/0$  and differentiate with respect to  $r$ , with  $x_i$  as constant parameters).

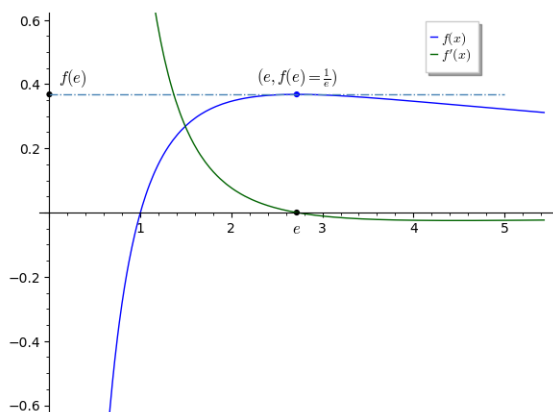
The following calculation uses the chain rule and knowledge of the derivative of the power function, must be read in reverse. The existence of the last limit implies the existence of the last-but-one, and so on.

In particular, when examining monotonicity and optimization problems (including local and absolute extrema), derivatives of the first and second orders are often sufficient to reveal the local behavior of a function defined on an interval of the real line.<sup>8</sup>

Notice in 5.3.2 we will revise the concepts of increasing and decreasing functions, thereby enriching our approach to studying the local behavior of functions. For further illustration, refer to the examples in 5.3.3. This latter aspect will be further elaborated upon in the initial section of Chapter 6. Therefore, the tasks presented below serve a more foundational purpose.

**5.C.15.** Discuss the monotonicity of the function  $f(x) = \frac{\ln(x)}{x}$  over its domain.

**Solution.** For any  $x$  in the domain  $A := (0, +\infty)$  of  $f$  we compute  $f'(x) = \frac{(\ln(x))'x - x'\ln(x)}{x^2} = \frac{1 - \ln(x)}{x^2}$ . Thus, the equation  $f'(x) = 0$  has a unique solution, namely  $x_0 = e$ . This is a “critical” or “stationary point” of  $f$ , see 5.3.2 for terminology. Since  $x^2$  is always positive, the sign of  $f'$  depends on the numerator  $1 - \ln(x)$ . Observe now that  $1 - \ln(x) > 0 \Leftrightarrow \ln(x) < 1 \Leftrightarrow \ln(x) < \ln(e) \Leftrightarrow x < e$ , and similarly we obtain  $1 - \ln(x) < 0 \Leftrightarrow x > e$ . Thus  $f'(x) > 0$  for all  $x \in (0, e)$  and  $f'(x) < 0$  for all  $x \in (e, +\infty)$ , which means that  $f$  is strictly increasing in the interval  $(0, e)$  and strictly decreasing in  $(e, +\infty)$ . In particular,  $f$  takes its maximum value at the point  $x_0$ , namely  $f(x_0) = f(e) = \frac{1}{e}$ , see also the figure below.



In section 6.1.2 we will learn about a test checking this final claim, which is based on the second derivative of  $f$ . The latter is simply defined by  $f'' = (f')'$ , hence in our case it has the form

$$f''(x) = (f'(x))' = \left(\frac{1 - \ln(x)}{x^2}\right)' = \frac{2 \ln(x) - 3}{x^3}.$$

<sup>8</sup>In this column we will mainly focus on real functions or on complex-valued functions of a real variable. Applications of complex-valued functions of a complex variable (complex analytic functions) will be analyzed later in Chapter 9.

$$\begin{aligned} \lim_{r \rightarrow 0} \ln(M^r(x_1, \dots, x_n)) &= \lim_{r \rightarrow 0} \frac{\ln\left(\frac{1}{n}(x_1^r + \dots + x_n^r)\right)}{r} \\ &= \lim_{r \rightarrow 0} \frac{x_1^r \ln x_1 + \dots + x_n^r \ln x_n}{\frac{x_1^r + \dots + x_n^r}{n}} \\ &= \frac{\ln x_1 + \dots + \ln x_n}{n} = \ln \sqrt[n]{x_1 \dots x_n}. \end{aligned}$$

Hence

$$\lim_{r \rightarrow 0} M^r(x_1, \dots, x_n) = \sqrt[n]{x_1 \dots x_n},$$

which is known as the *geometric mean*.

#### 4. Infinite sums and power series

The last part of this chapter is mainly devoted to infinite sums of numbers, aiming at infinite extension of polynomials – the so called power series.

We first complete the basic discussion of the exponential function, expressing it as a limit of polynomial approximations. This illustrates the more general need to develop effective tools to deal with sequences of numbers or functions. If the reader finds the next paragraphs too demanding, we suggest jumping to 5.4.3 starting the general discussion on infinite sums of numbers and maybe return later.

**5.4.1. The calculation of  $e^x$ .** For numerical computations, manipulation with limits of sequences is needed as well as addition and multiplication of scalars. Thus it might be a good idea to approximate non-polynomial functions by sequences of numbers that can be calculated easily, keeping control of the approximation errors.

We approach the function  $e^x$  this way. In view of the expected property  $(e^x)' = e^x$  (cf. 5.3.7), we look for a function whose rate of increase equals the function’s value at every point. This can be imagined as a splendid interest rate equal to the current value of your money.

If we apply the interest rate per year once a month, once a day, once an hour, and so on, we obtain the following values for the yield  $x$  of the deposit after one year:

$$\left(1 + \frac{x}{12}\right)^{12}, \quad \left(1 + \frac{x}{365}\right)^{365}, \quad \left(1 + \frac{x}{8760}\right)^{8760}, \dots$$

Therefore, we could guess that

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n.$$

At the same time, we can imagine that the finer we apply the interest, the higher the yield will be. So the sequence on the right-hand side should be an increasing sequence.

In detail, we examine the sequence of numbers

$$a_n = \left(1 + \frac{1}{n}\right)^n.$$

*Bernoulli’s inequality* will come in handy:

According to this criterium, whenever we have  $f''(x_0) < 0$  near a critical point  $x_0$  of  $f$ , then  $f$  must attain a (local) maximum at  $x_0$ . For our case we indeed compute  $f''(e) = -\frac{1}{e^3} < 0$ .

Notice finally in Sage one could compute the first and second derivative of  $f$  by the cell

```
f(x)=ln(x)/x; d1(x)=diff(f, x).factor()
show(d1(x))
D(x)=diff(d1, x).factor(); show(D(x))
```

Here we have programmed Sage to compute  $f''(x)$  “manually” (we denote this by  $D(x)$ ). An alternative approach for the second derivative relies on an application of the command `diff(f, x, 2).factor()`. This directly prints out the factorization of the expression given for the second derivative of  $f$ , without being necessary to compute the first derivative. We will meet more such applications in Chapter 6.  $\square$

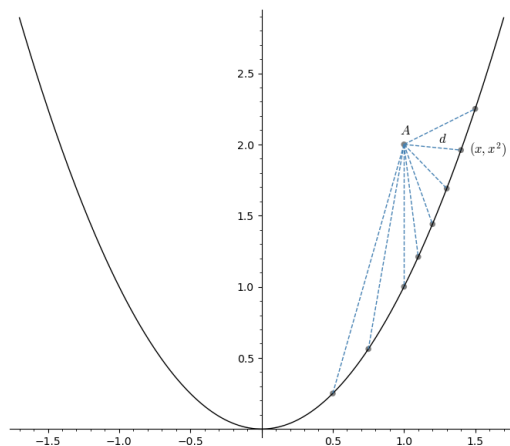
**5.C.16.** Prove by two different ways that the tangent function  $f(x) = \tan(x)$  is strictly increasing for all  $x \in (-\pi/2, \pi/2)$ .  $\circ$

**5.C.17.** Consider the parabola  $y = x^2$ . Determine the  $x$ -coordinate of the parabola’s point which is nearest to the point  $A$  with coordinates  $x_0 = 1$  and  $y_0 = 2$ .

**Solution.** Recall the formula for the Euclidean distance between two points  $[x_1, y_1], [x_2, y_2]$  in the plane:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

An arbitrary point in the parabola has coordinates  $[x, y] = [x, x^2]$  (and in this way we eliminate  $y$  from the problem). Since the task requires the minimization of the distance between the points  $A = [1, 2]$  and  $[x, x^2]$ , it is sufficient to find the minimum of the function  $d(x) = \sqrt{(x-1)^2 + (x^2-2)^2}$  with  $x \in \mathbb{R}$ , see also the figure for an illustration of the idea.



The function  $d(x)$  is non-zero for any  $x \in \mathbb{R}$ , in particular the equation  $g(x) := (x-1)^2 + (x^2-2)^2 = 0$  has only complex solutions. We have

$$d'(x) = \frac{g'(x)}{2\sqrt{g(x)}} = \frac{g'(x)}{2d(x)}, \quad x \in \mathbb{R}.$$

**Lemma.** For every real number  $b \geq -1$ ,  $b \neq 0$ , and a natural number  $n \geq 2$ ,  $(1+b)^n > 1+nb$ .

**PROOF.** For  $n = 2$ ,

$$(1+b)^2 = 1 + 2b + b^2 > 1 + 2b.$$

Proceed by induction on  $n$ , supposing  $b > -1$ . Assume that the proposition holds for some  $k \geq 2$  and calculate:

$$\begin{aligned} (1+b)^{k+1} &= (1+b)^k(1+b) > (1+kb)(1+b) \\ &= 1 + (k+1)b + kb^2 > 1 + (k+1)b \end{aligned}$$

The statement is, of course, true for  $b = -1$  as well.  $\square$

Now

$$\begin{aligned} \frac{a_n}{a_{n-1}} &= \frac{(1 + \frac{1}{n})^n}{(1 + \frac{1}{n-1})^{n-1}} = \frac{(n^2 - 1)^n n}{n^{2n}(n-1)} \\ &= \left(1 - \frac{1}{n^2}\right)^n \frac{n}{n-1} > \left(1 - \frac{1}{n}\right) \frac{n}{n-1} = 1. \end{aligned}$$

by using Bernoulli’s inequality with  $b = -\frac{1}{n^2}$ . So  $a_n > a_{n-1}$  for all natural numbers, and it follows that the sequence  $a_n$  is indeed increasing.

The following similar calculation (also using Bernoulli’s inequality) verifies that the sequence of numbers

$$b_n = \left(1 + \frac{1}{n}\right)^{n+1} = \left(1 + \frac{1}{n}\right) \left(1 + \frac{1}{n}\right)^n$$

is decreasing. Notice that  $b_n > a_n$ . Also,

$$\begin{aligned} \frac{b_n}{b_{n+1}} &= \frac{n}{n+1} \left(\frac{\frac{n+1}{n}}{\frac{n+2}{n+1}}\right)^{n+2} = \frac{n}{n+1} \left(\frac{n^2 + 2n + 1}{n^2 + 2n}\right)^{n+2} \\ &= \frac{n}{n+1} \left(1 + \frac{1}{n(n+2)}\right)^{n+2} \\ &\geq \frac{n}{n+1} \left(1 + \frac{n+2}{n(n+2)}\right) = 1. \end{aligned}$$

Thus the sequence  $a_n$  is increasing and bounded from above, so the set of its terms has a supremum which equals the limit of the sequence. At the same time, this value is also the limit of the decreasing sequence  $b_n$  because

$$\lim_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right) a_n = \lim_{n \rightarrow \infty} a_n.$$

This limit determines one of the most important numbers in mathematics (besides the numbers 0, 1, and  $\pi$ ), namely Euler’s number<sup>14</sup>  $e$ . Thus

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n.$$

<sup>14</sup>The ingenious Swiss mathematician, physicist, astronomer, logician and engineer Leonhard Euler (1707-1783) was behind extremely many inventions, including original mathematical techniques and tools.

Because the square root is an increasing function, the function  $d$  takes its least value at the same point where the function  $g$  does. The derivative of  $g$  has the form  $g'(x) = 4x^3 - 6x - 2$ , for all  $x \in \mathbb{R}$ . Moreover, the equation  $g'(x) = 0$  or equivalently  $0 = 2x^3 - 3x - 1$  has three solutions:

$$x_0 = 1, \quad x_1 = \frac{1 - \sqrt{3}}{2}, \quad x_2 = \frac{1 + \sqrt{3}}{2}.$$

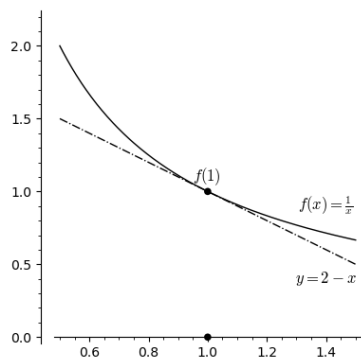
We deduce that the  $x$ -coordinate in question equals to  $x_2 = \frac{1 + \sqrt{3}}{2}$ . Can you explain how we decline  $x_0$  and  $x_1$ ?  $\square$

*First order derivatives and in particular tangent lines give us the ability to approximate functions locally by linear functions (recall that linear functions are the easiest functions to work with). In particular, when  $f$  is a differentiable function at a point  $x_0$  of its domain, we can approximate the values of  $f$  near  $x_0$  via the formula  $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$ . Here, the r.h.s is a first-degree polynomial with respect to  $x$  and we suppose that the difference  $x - x_0$  is "small enough". Hence using first-order derivatives we can obtain a tangent line approximation of the value  $f(x)$  in a small neighbourhood of  $x_0$ .*

*In Section A of Chapter 6 these ideas will be generalized. There we will explain how to approximate functions based on higher-degree polynomials (Taylor polynomials). Next we pose applications related to linear approximations, see also 6.A.37 and 6.A.38 for an alternative interpretation in terms of the so-called "differentials". More exercises on tangential approximations are presented in Section D.*

**5.C.18. Linear approximations.** Approximate linearly the function  $f(x) = 1/x$  near  $x_0 = 1$ , and present the graph of  $f$  and that of its tangent line at  $x_0$ . Then compute the approximations at  $x = 0.9$  and  $x = 1.1$  and compare their difference with the real values of  $f$  at these points.

**Solution.** The first derivative of  $f$  is given by  $f'(x) = -1/x^2$  and at  $x_0 = 1$  we get  $f'(x_0) = -1$ . Thus the tangent line of  $f$  passing from  $x_0$  has the form  $f(1) + (x - 1)f'(1) = 2 - x$ , that is, the linear approximation of  $f$  near  $x_0$  is given by the line  $L(x) = 2 - x$ . The figure below illustrate the situation:



**5.4.2. Power series for  $e^x$ .** The exponential function has been defined as the only continuous function satisfying  $f(1) = e$  and  $f(x + y) = f(x) \cdot f(y)$ . The base  $e$  is now expressed as the limit of the sequence  $a_n$ , thus necessarily

$$e^x = \lim_{n \rightarrow \infty} (a_n)^x.$$

Fix a real number  $x \neq 0$ . If we replace  $n$  with  $n/x$  in the numbers  $a_n$  from the previous paragraph, we arrive again at the same limit. (Think this out in detail!) Hence

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^{\frac{n}{x}}, \quad e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n.$$

Denote the  $n$ -th term of this sequence by  $u_n(x) = \left(1 + \frac{x}{n}\right)^n$  and express it by the binomial theorem:

$$\begin{aligned} u_n(x) &= 1 + n \frac{x}{n} + \frac{n(n-1)x^2}{2!n^2} + \dots + \frac{n!x^n}{n!n^n} \\ &= 1 + x + \frac{x^2}{2!} \left(1 - \frac{1}{n}\right) \\ &+ \frac{x^3}{3!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) + \dots \\ &+ \frac{x^n}{n!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{n-1}{n}\right). \end{aligned} \tag{1}$$

Look at  $u_n(x)$  for very large  $n$ . It seems that many of the first summands of  $u_n(x)$  will be fairly close to the values  $\frac{1}{k!}x^k$ ,  $k = 0, 1, \dots$ . Thus it is plausible that the numbers  $u_n(x)$  should be very close to  $v_n(x) = \sum_{j=0}^n \frac{1}{j!}x^j$  and thus both these sequences should have the same limit.

The following theorem is perhaps one of the most important results of Mathematics:

THE POWER SERIES FOR  $e^x$

**Theorem.** The exponential function  $e^x$  equals, for each  $x \in \mathbb{R}$ , the limit  $\lim_{k \rightarrow \infty} v_k(x)$  of the partial sums in the expression

$$e^x = 1 + x + \frac{1}{2}x^2 + \dots + \frac{1}{n!}x^n + \dots = \sum_{n=0}^{\infty} \frac{1}{n!}x^n.$$

The function  $e^x$  is differentiable at each point  $x$  and its derivative is  $(e^x)' = e^x$ .

**PROOF.** The technical proof makes the above idea precise. Although the complete argumentation might look complicated, the strategy is straightforward. We shall go through three steps.

- 1) Prove that the partial sums  $v_n$  converge.
- 2) Write  $u_{n,k}$  for the first  $k < n$  summands in  $u_n$  and conclude that for given  $k$ , the difference between  $v_k$  and  $u_{n,k}$  gets arbitrarily small for large  $n > k$ .
- 3) Show that there are subsequences  $v_{k_i}$  and  $u_{n_i}$  converging to the same limit. This will conclude the proof of the first claim of the theorem.



Now, we compute  $f(0, 9) \approx 1.1111$  and  $L(0, 9) = 1.1$ , thus  $|L(0.9) - f(0.9)| \approx 0.0111$ . Similarly, we have  $f(1.1) \approx 0.90909$ ,  $L(1.1) = 0.9$  and  $|L(1.1) - f(1.1)| \approx 0.009$ .  $\square$

**5.C.19.** Given that  $(\arcsin(x))' = \frac{1}{\sqrt{1-x^2}}$  for  $x \in (-1, 1)$ , approximate linearly the value  $\arcsin(0.497)$ .  $\circ$

**5.C.20.** Approximate linearly the values  $\sin\left(\frac{29\pi}{180}\right)$  and  $\sin\left(\frac{46\pi}{180}\right)$ .  $\circ$

*Rolle's theorem, mean value theorem and its generalizations (Cauchy's mean value theorem), are important theorems that highlight the significance of derivatives, see 5.3.8 and 5.3.9. Below we illustrate them quickly via examples, and further linked applications are presented in the final section of this chapter.*



**5.C.21.** Prove that the equation  $x^{2027} + 7x^3 - 5 = 0$  has a unique solution,

**Solution.** Consider the function  $f(x) = x^{2027} + 7x^3 - 5$ . Since  $f(0) = -5 < 0$  and  $f(1) = 3 > 0$  by the intermediate value theorem there exists some  $x_0 \in (0, 1)$  such that  $f(x_0) = 0$ . Suppose that  $x_1 \neq x_0$  is another (positive) root of  $f$ . Without loss of generality we may assume that  $x_0 < x_1$ . We have that  $f(x_0) = f(x_1)$ , and  $f$  is continuous on  $[x_0, x_1]$  and differentiable on  $(x_0, x_1)$ . Thus by Rolle's theorem the first derivative  $f'$  of  $f$  must have a root in  $(x_0, x_1)$ . However,  $f'(x) = 2027x^{2026} + 21x^2$  and thus we have  $f'(x) > 0$  for all  $x > 0$ , a contradiction. Thus  $f$  has only one zero in  $(0, +\infty)$ . On the other hand we see that  $f(x) < 0$  for all  $x \in (-\infty, 0]$ . You may verify this final claim in Sage, just by the cell

```
f(x)=x^(2027)+7*x^3-5
assume(x<0); bool(f(x)<0)
```

Thus  $f$  has a unique solution for all  $x \in \mathbb{R}$ .  $\square$

**5.C.22.** Using Rolle's theorem prove that the equation  $x^3 + px + q = 0$  with  $p, q \in \mathbb{R}$ ,  $p > 0$ , admits a unique real solution.

**Solution.** Consider the function  $f(x) = x^3 + px + q$  with  $x \in \mathbb{R}$ , where  $p, q \in \mathbb{R}$  and  $p > 0$ . It is easy to prove that  $\lim_{x \rightarrow -\infty} f(x) = -\infty$  and  $\lim_{x \rightarrow +\infty} f(x) = +\infty$ . Moreover,  $f$  takes both negative and positive values, thus there exists some  $\xi \in \mathbb{R}$  with  $f(\xi) = 0$ . We will show that  $\xi$  is unique. Indeed, suppose that  $\zeta \in \mathbb{R}$  is another real number satisfying  $f(\zeta) = 0$ . We may assume that  $\zeta < \xi$ , and similarly is treated the case with  $\xi < \zeta$ . Based on the Rolle's theorem we can then find some  $a \in (\zeta, \xi)$  with  $f'(a) = 0$ . However,  $f'(x) = 3x^2 + p$  for all  $x$ , and so this gives a contradiction (since  $p > 0$ , the equation  $f'(x) = 0$  does not admit a real solution).  $\square$

**5.C.23.** Decide if the function  $f(x) = x^3 + b$ , where  $b$  is some constant, satisfies the mean value theorem on the interval  $[1, 2]$ . In the positive case determine  $c$ , where  $c$  is as in the Theorem 5.3.9.

1) Fix  $x$  and recall that  $v_n(x)$  is the sequence defined as the sums of the first  $n$  terms of the formal infinite expression

$$\sum_{j=0}^{\infty} c_j = \sum_{j=0}^{\infty} \frac{1}{j!} x^j.$$

For all  $m > n$ ,

$$|v_m(x) - v_n(x)| \leq \sum_{k=n+1}^m \frac{1}{k!} |x|^k = v_m(|x|) - v_n(|x|).$$

Consequently, in order to prove that  $v_n(x)$  is a Cauchy sequence, and thus convergent, it is enough to restrict ourselves to  $x > 0$  and to prove that  $v_n(x)$  is always bounded and thus convergent (as a nondecreasing sequence) in this case.

The quotient of adjacent terms in the series is  $\frac{c_{j+1}}{c_j} = \frac{x}{n+1}$ . Thus for every fixed  $x > 0$ , there is a number  $N \in \mathbb{N}$  such that  $\left| \frac{c_{j+1}}{c_j} \right| < \frac{1}{2}$  for all  $j \geq N$ . However, such large indices  $j$  satisfy  $|c_{j+1}| < \frac{1}{2}|c_j| < 2^{-(j-N+1)}|c_N|$ .

Recall that sums of a geometric series is computed from the equality  $(1 - q)(1 + q + \dots + q^k) = 1 - q^{k+1}$ . This means that the partial sums of the first  $n > N$  terms of our formal sum with  $x > 0$  can be estimated as follows

$$|v_n(x) - \sum_{j=0}^{N-1} \frac{1}{j!} x^j| < \frac{1}{N!} x^N \sum_{j=0}^{n-N} \frac{1}{2^j}.$$

In particular, the limit of the expressions on the right-hand side for  $n$  approaching infinity surely exists, and so the limit of the increasing sequence  $v_n$  also exists.

2) Consider some fixed  $k$  and  $\varepsilon > 0$ . If  $N$  is large enough, then clearly for all  $n > N$ ,  $|u_{n,k} - v_k| < \varepsilon$ . Indeed, there is only a fixed number of the brackets in the  $k$  summands of  $u_{n,k}$ , see (1), and they will all be arbitrarily close to 1, if  $n$  is large enough.

3) Next, consider  $x > 0$  and notice

$$|u_n - u_{n,k}| < |v_n - v_k|.$$

Thus, if we fix some  $\varepsilon > 0$ , then there is some  $M$ , such that for  $k > M$  we ensure  $|u_n - u_{n,k}| < \varepsilon$ , for all  $n > k$ . If  $x < 0$ , we may first bound the left hand side by sum of absolute values of the individual terms, which is still less than  $|v_n - v_k|$  evaluated for  $|x|$ .

Now, take such a  $k > M$  and consider  $N$  from the previous step (of course dependent on  $k$ ). Then for  $n > N$  we arrive at

$$|v_k - u_n| \leq |v_k - u_{n,k}| + |u_{n,k} - u_n| < 2\varepsilon.$$

Finally, choosing  $\varepsilon_\ell = \frac{1}{2\ell}$ , the previous estimate allows to find subsequences  $v_{k_\ell}, u_{n_\ell}$  satisfying

$$|v_{k_\ell} - u_{n_\ell}| < \frac{1}{\ell}.$$

Thus, the convergent sequences  $u_n$  and  $v_n$  must converge to the same limit value. This is the first claim we had to prove.

Remind we already know that  $(e^x)' = e^x$  if and only if the derivative equals 1 at the origin, see 5.3.7. Thus, it

**Solution.** The given function is a polynomial, and hence continuous on its domain  $[1, 2]$ . It is also differentiable on the open interval  $(1, 2)$ , with  $f'(x) = 3x^2$ . Therefore, it satisfies the mean value theorem. In particular, by Theorem 5.3.9 there exists  $c \in (1, 2)$  such that  $f'(c) = 3c^2 = \frac{f(2)-f(1)}{2-1} = 8+b-1-b = 7$ . Thus  $c^2 = 7/3$ , that is,  $c = \pm\sqrt{7/3}$ . Since  $-\sqrt{7/3}$  does not belong on the interval  $[1, 2]$ , we deduce that  $c = \sqrt{7/3}$  is the only possible value.  $\square$

**5.C.24.** Check which of the following functions satisfy the mean value theorem. In the positive case, find the possible values of  $c$ , where  $c$  is as in the Theorem 5.3.9:

$$f(x) = x^{\frac{2}{3}}, \quad g(x) = \frac{x}{x+1}, \quad h(x) = \sqrt[3]{8-x}$$

for  $x \in [-2, 2]$ ,  $x \in [1, 4]$  and  $x \in [0, 8]$ , respectively.  $\circ$

**5.C.25.** Using the mean value theorem prove that

$$-1 \leq \cos(2x) < \frac{\sin(2x)}{2x} \leq 1, \quad \text{for all } x \in \left(0, \frac{\pi}{2}\right).$$

**Solution.** For  $x = \frac{\pi}{2}$  the given inequality becomes  $-1 \leq -1 < 0 \leq 1$ , which is true. Hence we may focus on  $x \in (0, \frac{\pi}{2})$ . For any  $x \in (0, \frac{\pi}{2})$  consider the function

$$f(a) = \sin(2a), \quad \text{with } a \in [0, x].$$

Observe that  $f$  is continuous on  $[0, x]$  as a composition of continuous functions. It is also differentiable on the open interval  $(0, x)$  with  $f'(a) = 2\cos(2a)$ . Thus, by the mean value theorem there exists  $c \in (0, x)$  with  $f'(c) = \frac{f(x)-f(0)}{x-0}$ , or equivalently

$$2\cos(2c) = \frac{\sin(2x)}{x} \Leftrightarrow \cos(2c) = \frac{\sin(2x)}{2x}. \quad (*)$$

Next we may assume that  $0 < c < x < \frac{\pi}{2}$ , that is,  $0 < 2c < 2x < \pi$ . Also, for the function  $g(x) = \cos(x)$  we have  $g'(x) = -\sin(x) < 0$  for all  $x \in (0, \pi)$ . This means that the function  $\cos$  is strictly decreasing on  $(0, \pi)$  (having a local maximum at  $x = 0$  and a local minimum at  $x = \pi$ ). Thus, the previous inequality implies that

$$1 = \cos(0) \geq \cos(2c) > \cos(2x) \geq \cos(\pi) = -1.$$

Combining this inequality with  $(*)$ , the claim follows.  $\square$

**5.C.26.** Based on Cauchy's mean value theorem, show that

$$a < \frac{n}{n+1} \left( \frac{b^{n+1} - a^{n+1}}{b^n - a^n} \right) < b$$

for all  $n \in \mathbb{N}^*$ , where  $a, b \in \mathbb{R}$  are such that  $0 < a < b$   $\circ$

*The theory of derivatives is useful in a plethora of practical tasks, but also for crucial problems appearing in many other sciences and technology, e.g., in physics, chemistry, statistics, economics, biology, engineering, environmental science, informatics and computer science, architecture, science of graphics, and space science, to mention a few of them.*

*But why derivatives can be so useful? Partially this is because the derivative  $\frac{dy}{dx}$  represents the rate of change of  $y$*



remains to compute

$$\lim_{x \rightarrow 0} \frac{(1 + x + \frac{1}{2}x^2 + \dots) - 1}{x} = \lim_{x \rightarrow 0} \frac{x + \frac{1}{2}x^2 + \dots}{x}.$$

This seems to be tricky, since there are two limit expressions there (notice that an  $x$  may be cancelled since this is a constant in the inner limit):

$$\lim_{x \rightarrow 0} \left( \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{1}{k!} x^{k-1} \right) = 1 + \lim_{x \rightarrow 0} \left( \lim_{n \rightarrow \infty} \sum_{k=2}^n \frac{1}{k!} x^{k-1} \right).$$

But now, for each positive  $\varepsilon > 0$  we can find  $N \in \mathbb{N}$  such that

$$\lim_{n \rightarrow \infty} \left| \sum_{k=N}^n \frac{1}{(k+1)!} x^k \right| < \lim_{n \rightarrow \infty} \sum_{k=N}^n \frac{1}{(k+1)!} < \varepsilon$$

for all  $x \in [-1, 1]$ . Finally, we can restrict the interval for  $x$  enough to ensure that the remaining first terms are bounded by  $\varepsilon$ , too:

$$\sum_{k=2}^{N-1} \frac{1}{(k+1)!} x^k < \varepsilon.$$

This shows that the limit expression on the right-hand side must be zero. Thus the derivative exists and equals to one, as expected.  $\square$



Readers who skipped the preceding paragraphs (it doesn't matter whether on purpose or in need) can stay calm – we deduce all the results on the exponential function later again, using more general tools. In particular, we will see that all power series are always differentiable and can be differentiated term by term. We see later that the conditions  $f'(x) = f(x)$  and  $f(0) = 1$  determine a function uniquely.

**5.4.3. Number series.** When deriving the previous theorems about the function  $e^x$ , we have automatically used several extraordinarily useful concepts and tools. Now, we come back to them in detail:

#### INFINITE NUMBER SERIES

**Definition.** An *infinite series of numbers* is an expression

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \dots + a_k + \dots,$$

where the  $a_n$ 's are real or complex numbers. The sequence of *partial sums* is given by the terms  $s_k = \sum_{n=0}^k a_n$ . The series converges and equals  $s$  if the limit

$$s = \lim_{k \rightarrow \infty} s_k$$

of the partial sums exists and is finite.

If the sequence of partial sums has an improper limit, the series *diverges* to  $\infty$  or  $-\infty$ . If the limit of the partial sums does not exist, the series *oscillates*.

An immediate consequence of the properties of limits is the following claim:

with respect to  $x$ . Thus, the rate of change of any physical quantity at any time is obtained by differentiating the physical quantity with respect to time. Another reason is the related theory of minimal/maximals (optimization problems) and we will meet many such tasks in Chapter 6. The next few examples aim to present how derivatives may involve in problems of every-day life and other nice applications. An enrichment of this list, which roughly speaking can be long enough, is postponed to the final section E of this chapter.

**5.C.27.** For a company renting city taxis, any driver costs €6 per working hour (h), while the gas costs €1.40 per litter (ℓ). Suppose that inside the city these taxis can succeed  $(180 - 2u)/6$  kilometres (km) per litter of gas, where  $u$  denotes the driving velocity, enumerated in km/h. Find the cheapest driving speed so that the company can minimize its costs per kilometre (corresponding only to a taxi of its property), and present the corresponding minimal cost value in €/km per taxi.



**Solution.** We can focus on the cost corresponding to the use of a single taxi (in € per km), which should minimize. This cost is given by the sum of the costs of the used driver (per km) and of the used gas (per km). According to the statement we have:

- The driver costs  $\frac{6 \text{ €/h}}{u \text{ km/h}} = \frac{6}{u} \text{ €/km}$ ;
- The gas costs  $\frac{1.40 \text{ €/ℓ}}{\frac{(180-2u)}{6} \text{ km/ℓ}} = \frac{8.4}{180-2u} \text{ €/km}$ .

Thus, the total cost in € per km is a function with respect to  $u$ , given by

$$c(u) := \frac{6}{u} + \frac{8.4}{180-2u} = \frac{6}{u} - \frac{4.2}{u-90}.$$

First we see that the driving velocity  $u$ , should satisfy  $0 < u < 90$  (e.g., the cost blows up for  $u = 90$ ), and for such  $u$  the extreme points of  $c(u)$  correspond to the solutions of the equation  $c'(u) = 0$  (we recommend to sketch the graph of the cost function  $c$  via Sage). The derivative of  $c$  is given by

$$c'(u) = -\frac{6}{u^2} + \frac{4.2}{(u-90)^2}$$

and it follows that the equation  $c'(u) = 0$  has two solutions, namely  $u_{1,2} = \pm 30\sqrt{70} + 300$  km/h. These computations can be quickly verified in Sage:

```
u=var("u"); c(u)=(6/u)+8.4/(180-2*u)
dc(u)=diff(c(u), u); show(solve(dc==0, u))
```

Acceptable is only the value  $u_1 = -30\sqrt{70} + 300 \approx 49$  km/h, and this is the speed with the lowest costs. In particular, with this velocity the cost value  $c(u_1)$  of a single taxi is approximately 0.225 €/km. □

**5.C.28.** Suppose that from a given piece of industrial paper having area  $\alpha^2$  we want to construct an open box, by cutting

**Corollary.** Consider two convergent series  $S = \sum_{n=0}^{\infty} a_n$  and  $T = \sum_{n=0}^{\infty} b_n$ . Then their sum and constant multiple by real or complex number  $c$  are convergent too, and

$$S + T = \sum_{n=0}^{\infty} (a_n + b_n), \quad cS = \sum_{n=0}^{\infty} ca_n.$$

In particular, any linear combination of series with constant coefficients is convergent and equals to the same linear combination of the sums. Check the details yourself!

**5.4.4. Properties of series.** For the sequence of partial sums  $s_n$  to converge, it is necessary and sufficient that it is a Cauchy sequence; that is

$$|s_m - s_n| = |a_{n+1} + \dots + a_m|$$

must be arbitrarily small for sufficiently large  $m > n$ . Since

$$|a_{n+1}| + \dots + |a_m| \geq |a_{n+1} + \dots + a_m|,$$

the convergence of the series  $\sum_{k=0}^{\infty} |a_k|$  implies the convergence of the series  $\sum_{k=0}^{\infty} a_k$ .

#### ABSOLUTELY CONVERGENT SERIES

A series  $\sum_{k=0}^{\infty} a_k$  is *absolutely convergent* if the series  $\sum_{k=0}^{\infty} |a_k|$  converges.



Absolute convergence is introduced because it is often much easier verified. The following theorem shows that all simple algebraic operations behave “very well” for series that converge absolutely.

#### PROPERTIES OF SERIES

**Theorem.** Let  $S = \sum_{n=0}^{\infty} a_n$  and  $T = \sum_{n=0}^{\infty} b_n$  be two absolutely convergent series. Then

(1) their sum converges absolutely to the sum

$$S + T = \sum_{n=0}^{\infty} a_n + \sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} (a_n + b_n),$$

(2) their difference converges absolutely to the difference

$$S - T = \sum_{n=0}^{\infty} a_n - \sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} (a_n - b_n),$$

(3) their product converges absolutely to the product

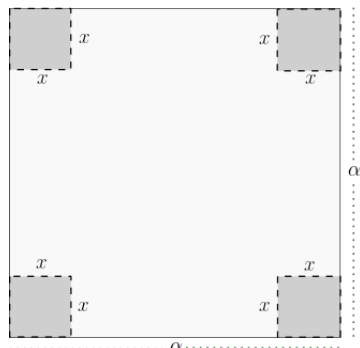
$$S \cdot T = \left( \sum_{n=0}^{\infty} a_n \right) \cdot \left( \sum_{n=0}^{\infty} b_n \right) = \sum_{n=0}^{\infty} \left( \sum_{k=0}^n a_{n-k} b_k \right),$$

(4) the value  $S$  of the sum does not depend on any rearrangement of the series, i.e.,  $\sum_{n=0}^{\infty} a_{\sigma(n)} = S$  for any permutation  $\sigma : \mathbb{N} \rightarrow \mathbb{N}$  of integers.

**PROOF.** The convergence in the first and the second statements were discussed already. Further, the linear combinations satisfy  $|\alpha a_n + \beta b_n| \leq |\alpha| |a_n| + |\beta| |b_n|$  and thus the absolute convergence is obvious, too.

four small squares of equal size of its corners and then bending the resulting sides. Determine the total area of paper that we should cut, so that the box has maximum volume.

**Solution.** Let  $x$  be the length of the side of any of the squares that we should cut, as it is illustrated in the figure:



The volume of the box under construction is given by  $V(x) = x(\alpha - 2x)^2$  with  $0 < x < \frac{\alpha}{2}$ . We want to locate those  $x$  for which the function  $V$  attains its maximum value  $V(x)$ . Since  $\alpha > 0$  is a constant, an application of the product rule gives

$$\begin{aligned} V'(x) &= \frac{dV}{dx} = (\alpha - 2x)^2 + 2 \cdot x \cdot (\alpha - 2x) \cdot (-2) \\ &= (\alpha - 2x)(\alpha - 6x). \end{aligned}$$

Thus  $V'(x) = 0$  if and only if  $x = \frac{\alpha}{2}$  or  $x = \frac{\alpha}{6}$ . Now we will use the theory of second derivatives, which are analyzed in Section 6.1.5. As we said above, the second derivative of a differentiable function  $f$  is the derivative of its first derivative. So

$$\begin{aligned} V''(x) &= (V'(x))' = -2 \cdot (\alpha - 6x) - 6(\alpha - 2x) \\ &= 24x - 8\alpha = 24 \left( x - \frac{\alpha}{3} \right). \end{aligned}$$

According to 6.1.5, if  $x_0$  satisfies  $f'(x_0) = 0$  and  $f''(x_0) < 0$ , then  $x_0$  is a local maximum of  $f$ , while when  $x_0$  is such that  $f'(x_0) = 0$  but  $f''(x_0) > 0$ , then  $x_0$  is a local minimum of  $f$ . In our case we see that

$$V''\left(\frac{\alpha}{2}\right) = 4\alpha > 0, \quad V''\left(\frac{\alpha}{6}\right) = -4\alpha < 0.$$

Thus, the volume of the box under construction will be maximal if and only if  $x = \frac{\alpha}{6}$ , so the area that we should cut is  $4 \cdot \frac{\alpha^2}{36} = \frac{\alpha^2}{9} = \left(\frac{\alpha}{3}\right)^2$ . Moreover, the maximal volume is given by  $V\left(\frac{\alpha}{6}\right) = \frac{2\alpha^3}{27} = 2\left(\frac{\alpha}{3}\right)^3$ .

An implementation of the whole problem via Sage is easy, and can be done in many ways. For instance, type the cell

```
var("a") #introduce a symbolic variable
V(x)=x*(a-2*x)^2 #declare the volume
DV(x)=diff(V, x).factor() #declare the 1st
#derivative of the volume
show(solve(DV(x)==0, x)) #critical point eq.
show(diff(V, x, 2)(x=a/6).factor())
show(V(a/6).factor()) #the maximal volume
```

The third statement is not so simple. Write

$$c_n = \sum_{k=0}^n a_{n-k} b_k.$$

From the assumptions and from the rule for the limit of a product,

$$\left( \sum_{n=0}^k a_n \right) \cdot \left( \sum_{n=0}^k b_n \right) \rightarrow \left( \sum_{n=0}^{\infty} a_n \right) \cdot \left( \sum_{n=0}^{\infty} b_n \right).$$

Thus it suffices to prove that

$$0 = \lim_{k \rightarrow \infty} \left( \left( \sum_{n=0}^k a_n \right) \cdot \left( \sum_{n=0}^k b_n \right) - \sum_{n=0}^k c_n \right).$$

Consider the expressions

$$\begin{aligned} \left( \sum_{n=0}^k a_n \right) \cdot \left( \sum_{n=0}^k b_n \right) &= \sum_{0 \leq i, j \leq k} a_i b_j, \\ c_n &= \sum_{\substack{i+j=n \\ 0 \leq i, j \leq k}} a_i b_j, \quad \sum_{n=0}^k c_n = \sum_{\substack{i+j \leq k \\ 0 \leq i, j \leq k}} a_i b_j. \end{aligned}$$

along with the bound

$$\begin{aligned} \left| \left( \sum_{n=0}^k a_n \right) \cdot \left( \sum_{n=0}^k b_n \right) - \sum_{n=0}^k c_n \right| &= \left| \sum_{\substack{i+j > k \\ 0 \leq i, j \leq k}} a_i b_j \right| \\ &\leq \sum_{\substack{i+j > k \\ 0 \leq i, j \leq k}} |a_i b_j|. \end{aligned}$$

If the sum of the indices is to be larger than  $k$ , then at least one of them must be larger than  $k/2$ . The expression does not decrease if more terms are added into it. Take all as in the product and remove only those whose indices are both at most  $k/2$ .

$$\sum_{\substack{i+j > k \\ 0 \leq i, j \leq k}} |a_i b_j| \leq \sum_{0 \leq i, j \leq k} |a_i b_j| - \sum_{0 \leq i, j \leq k/2} |a_i b_j|.$$

However, both the expressions of the difference are the partial sums for the product  $S \cdot T$ . Therefore, they share the same limit and their difference goes to zero.

The last claim seems to be a little tricky. Notice, for each small  $\varepsilon > 0$  we can find a common bound  $\alpha$  such that for all  $N > \alpha$  both estimates are true:

$$\sum_{n=N}^{\infty} |a_n| < \varepsilon, \quad \left| \sum_{n=0}^N a_n - S \right| < \varepsilon.$$

Now, consider any permutation  $\sigma$  of the indices and write  $\square I_\sigma = \{\sigma^{-1}(0), \dots, \sigma^{-1}(\alpha)\}$ . Then, for each  $N > \max I_\sigma$

**5.C.29.** On a spherical balloon we inflate air so that its volume increase with a rate of  $16 \text{ cm}^3/\text{min}$ . Find the rate of change of the radius of the balloon when its volume is  $36 \text{ cm}^3$ . Hint: The volume of a sphere of radius  $r$  is given by  $\frac{4\pi r^3}{3}$ .

**Solution.** The radius of the balloon can be viewed as a function with respect to the time  $t$ . Hence also the volume of the balloon is a function of  $t$  (since it depends on the radius), i.e.,  $V(t) = \frac{4\pi}{3}r^3(t)$  for all  $t$ .

Now, by assumption at certain time  $t = t_0$  the volume is  $36 \text{ cm}^3$ . Hence we have  $36 = \frac{4\pi r^3(t_0)}{3}$ , and thus  $r^3(t_0) = \frac{3^3}{\pi}$ , that is,  $r(t_0) = \frac{3}{\sqrt[3]{\pi}}$ . We also compute

$$V'(t) = \frac{dV}{dt} = 4\pi r^2(t)r'(t).$$

By assumption  $\frac{dV}{dt} = 16$ , hence  $16 = 4\pi r^2(t)r'(t)$  and this gives  $r'(t) = \frac{dr}{dt} = \frac{4}{\pi r^2(t)}$ . At  $t = t_0$  we have  $r'(t_0) = \frac{4}{\pi r^2(t_0)}$  and by replacing the value  $r(t_0) = \frac{3}{\sqrt[3]{\pi}}$  we obtain the rate of change of the radius of the balloon at time  $t = t_0$ .  $\square$

**5.C.30.** Consider an isosceles triangle with base length  $b$  and height (above the base)  $h$ . Inscribe the rectangle having the greatest possible area into it (one of the rectangle's sides will lie on the triangle's base). Determine the area  $S$  of this rectangle.

**Solution.** To solve this problem, it suffices to consider the problem of inscribing the largest rectangle into a right triangle with legs of lengths  $b/2$  and  $h$  so that two of the rectangle's sides lie on the legs of the triangle. Thus we reduce the problem to maximizing the function

$$f(x) = x \left( h - \frac{2hx}{b} \right)$$

on the closed interval  $I = [0, b/2]$ . Observe that  $f(x) \geq 0$ , for all  $x \in I$ , with  $f(0) = f(b/2) = 0$ . We also have  $f'(x) = h - \frac{4hx}{b}$  and  $x_0 = b/4$  is the unique stationary point of  $f|_{[0, b/2]}$ . There  $f$  takes its greatest value, hence the sides of the required rectangle are  $b/2$  long (i.e., twice  $x_0$ ), and  $h/2$  long (the latter is obtained by substituting  $b/4$  for  $x$  into the expression  $h - 2hx/b$ ). Moreover,  $S = hb/4$ .  $\square$

**5.C.31.** Recall that the velocity of a moving object is the derivative of its position function, and its acceleration is the derivative of its velocity function. If the position of a moving object in time  $t$  is given by the function

$$s(t) = -(t - 3)^2 + 16, \quad t \in [0, 7],$$

where the time  $t$  measured in seconds, and the position is measured in meters, determine the following:

- (a) the initial velocity of the object (that is, at  $t = 0$ );
- (b) the time and position at which its velocity is zero;
- (c) its velocity and acceleration at time  $t = 4$  s.  $\circ$

In the reminder of this section we shall exercise the use of the so called "l'Hopital's rule" via a variety of examples (for extra material refer to Section E in the end of this chapter).



clearly

$$\begin{aligned} \left| \sum_{n=0}^N a_{\sigma(n)} - S \right| &= \left| \sum_{n \leq N, n \in I_\sigma} a_{\sigma(n)} - S + \sum_{0 \leq n \leq N, n \notin I_\sigma} a_{\sigma(n)} \right| \\ &\leq \left| \sum_{n=0}^\alpha a_n - S \right| + \sum_{0 \leq n \leq N, n \notin I_\sigma} |a_{\sigma(n)}|. \end{aligned}$$

Next, notice that  $n \notin I_\sigma$  means  $\sigma(n) > \alpha$ . Thus, the latter term is at most equal to  $\sum_{\alpha+1}^\infty a_n$  and thus the entire expression is bounded by  $2\varepsilon$ . This shows that the rearranged series converges to the same value  $S$  again.  $\square$

**5.4.5. Simple tests.** The following theorem collects some useful conditions for deciding on the convergence of series.

#### CONVERGENCE TESTS

**Theorem.** Let  $S = \sum_{n=0}^\infty a_n$  be an infinite series of real or complex numbers. Let  $T = \sum_{n=0}^\infty b_n$  be another series with all  $b_n \geq 0$  real.

- (1) If the series  $S$  converges, then  $\lim_{n \rightarrow \infty} a_n = 0$ .
- (2) (The comparison test) If  $T$  converges and  $|a_n| \leq b_n$ , then  $S$  converges absolutely. If  $b_n \leq |a_n|$  and  $T$  diverges, then  $S$  does not converge absolutely.
- (3) (The limit comparison test). If both  $a_n$  and  $b_n$  are positive real numbers and the finite limit

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = r > 0$$

exists, then  $S$  converges if and only if  $T$  converges.

- (4) (The ratio test) Suppose that the limit of the quotients of adjacent terms of the series exists and

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = q.$$

Then the series  $S$  converges absolutely for  $|q| < 1$  and does not converge absolutely for  $|q| > 1$ . If  $|q| = 1$  the series may or may not converge.

- (5) (The root test) If the limit

$$\lim_{n \rightarrow \infty} \sqrt[n]{|a_n|} = q$$

exists, then the series converges absolutely for  $q < 1$ . It does not converge absolutely for  $q > 1$ . If  $q = 1$ , the series may or may not converge.



**PROOF.** (1) The existence and the potential value of the limit of a sequence of complex numbers is given by the limits of the real parts and the imaginary parts. Thus it suffices to prove the first proposition for sequences of real numbers. If  $\lim_{n \rightarrow \infty} a_n$  does not exist or is non-zero, then for a sufficiently small number  $\varepsilon > 0$ , there are infinitely many terms  $a_k$  with  $|a_k| > \varepsilon$ . There are either infinitely many positive terms or infinitely many negative terms among them. But then, adding any one of them into the partial sum, the difference of the adjacent terms  $s_n$  and  $s_{n+1}$  is at least

Recall the claim is that for each differentiable curve  $[g(t), f(t)]$  emanating from the origin, the limit of the slope  $\frac{f(t)}{g(t)}$  exists and they are equal, provided that the limit of the tangent lines in the origin exists. This resolves the indeterminate forms  $\frac{0}{0}$ , and the similar result is available for indeterminate forms  $\frac{\infty}{\infty}$ , see also 5.3.10.

It's important to note that multiple applications of l'Hopital's rule are quite common. This practice involves stating that the limits of quotients of derivatives obtained through l'Hopital's rule are equal to the limits of the original quotients, albeit this is sometimes an abuse of notation. To justify this, it is essential to ensure that the limits produced on the right-hand sides actually exist, which may require repeated verification.

**5.C.32. l'Hopital's rule.** Apply the l'Hopital's rule to compute the limit  $\lim_{x \rightarrow +\infty} f(x)$ , where

$$f(x) = \frac{\ln x - 4e^x}{8e^x \alpha^e - \ln x}, \quad \alpha = \text{constant} > 0.$$

**Solution.** When  $x \rightarrow +\infty$ , both the numerator and denominator of  $f(x)$  give the indeterminate form  $(+\infty) - (+\infty)$ . However, we may write

$$\begin{aligned} \lim_{x \rightarrow +\infty} f(x) &= \lim_{x \rightarrow +\infty} \frac{\ln x - 4e^x}{8e^x \alpha^e - \ln x} = \lim_{x \rightarrow +\infty} \frac{\frac{\ln x}{e^x} - 4}{8\alpha^e - \frac{\ln x}{e^x}} \\ &= \frac{\lim_{x \rightarrow +\infty} \frac{\ln x}{e^x} - 4}{8\alpha^e - \lim_{x \rightarrow +\infty} \frac{\ln x}{e^x}}. \end{aligned}$$

Now, for the limit  $\lim_{x \rightarrow +\infty} \frac{\ln x}{e^x}$  we may apply the l'Hopital's rule. For this we need the derivatives of  $\ln x$  and  $e^x$ , and by 5.3.7 we know that  $(\ln x)' = 1/x$  and  $(e^x)' = e^x$ , respectively. Therefore

$$\begin{aligned} \lim_{x \rightarrow +\infty} f(x) &= \frac{\lim_{x \rightarrow +\infty} \frac{(\ln x)'}{(e^x)'} - 4}{8\alpha^e - \lim_{x \rightarrow +\infty} \frac{(\ln x)'}{(e^x)'}} = \frac{\lim_{x \rightarrow +\infty} \frac{1/x}{e^x} - 4}{8\alpha^e - \lim_{x \rightarrow +\infty} \frac{1/x}{e^x}} \\ &= \frac{\lim_{x \rightarrow +\infty} \frac{1}{x e^x} - 4}{8\alpha^e - \lim_{x \rightarrow +\infty} \frac{1}{x e^x}} = \frac{0 - 4}{8\alpha^e - 0} = -\frac{1}{2\alpha^e}. \end{aligned}$$

**5.C.33.** Based on the l'Hopital's rule verify the given type and next verify the given result:

- (a)  $\lim_{x \rightarrow 0} \frac{\sin(2x) - 2 \sin(x)}{2e^x - x^2 - 2x - 2} = -3$ , (type  $\frac{0}{0}$ );
- (b)  $\lim_{x \rightarrow 0^+} \frac{\ln(x)}{\cot(x)} = 0$ , (type  $\frac{\infty}{\infty}$ );
- (c)  $\lim_{x \rightarrow 1^+} \left( \frac{x}{x-1} - \frac{1}{\ln(x)} \right) = \frac{1}{2}$ , (type  $\infty - \infty$ );

$\varepsilon$ . Thus the sequence of partial sums cannot be a Cauchy sequence and, therefore, it cannot be convergent.

(2) We are dealing with absolute convergence and thus the sequence of partial sums is non-decreasing. Once bounded, such a sequence is convergent and converges to its supremum. Similarly, the sequence of partial sums must be unbounded, if estimated from below by a divergent sequence.

(3) Since limit  $r = \lim_{n \rightarrow \infty} \frac{a_n}{b_n}$  exists, for any given  $\varepsilon > 0$  and sufficiently big  $n > N_\varepsilon$ ,

$$(r - \varepsilon)b_n < a_n < (r + \varepsilon)b_n.$$

Thus, after choosing  $\varepsilon < r$  it follows that  $a_n < (r + \varepsilon)b_n$  and  $b_n < \frac{1}{r - \varepsilon}a_n$ . The result follows from the previous claim (2).

(4) To prove absolute convergence, it can be assumed that the terms of the series are real numbers  $a_n > 0$ . Suppose  $q < r < 1$  for a real number  $r$ . From the existence of the limit of the quotients, for every  $j$  greater than a sufficiently large  $N$ ,

$$a_{j+1} < r \cdot a_j \leq r^{(j-N+1)} a_N,$$

But this means that the partial sums  $s_n$  are, for large  $n > N$ , bounded from above by the sums

$$s_n < \sum_{j=0}^N a_j + a_N \sum_{j=0}^{n-N} r^j = \sum_{j=0}^N a_j + a_N \frac{1 - r^{n-N+1}}{1 - r}$$

where the last equality follows from the general equality  $(1 - r)(1 + r^2 + \dots + r^k) = 1 - r^{k+1}$ . Since  $0 < r < 1$ , the set of all partial sums is an increasing sequence bounded from above, and thus its limits equals to its supremum.

In the case  $q > r > 1$ , a similar technique can be used. However, this time, from the existence of the limit of the quotients,

$$a_{j+1} > r \cdot a_j \geq r^{(j-N+1)} a_N > 0.$$

This implies that the absolute values of the particular terms of the series do not converge to zero, and thus the series cannot be convergent, by the already proved part (1) of the theorem.

(5) The proof is similar to the previous case. From the existence of the limit  $q < 1$ , it follows that for any  $r, q < r < 1$ , there is an  $N$  such that for all  $n > N$ ,  $\sqrt[n]{|a_n|} < r$  holds. Exponentiation then gives  $|a_n| < r^n$ , there is a comparison with a geometric series. Thus the proof can be finished in the same way as in the case of the ratio test.  $\square$

**5.4.6. Limes superior and inferior.** In the proofs of the last two statements of the previous theorem, a much weaker assumption is used than the existence of the limit. It is only necessary to know that the examined sequences of non-negative terms (ratios or roots) are, from a given index on, either all larger or all less than a given number.

For this purpose however, it suffices to consider, for a given bounded sequence of terms  $b_n$ , the supremum of the terms with index higher than  $n$ . These suprema always exist and create a non-increasing sequence. Its infimum is then



(d)  $\lim_{x \rightarrow 1^+} (\ln(x) \cdot \ln(x-1)) = 0$ , (type  $0 \cdot \infty$ );

(e)  $\lim_{x \rightarrow 0^+} (\cot(x))^{\frac{1}{\ln(x)}} = \frac{1}{e}$ , (type  $\infty^0$ );

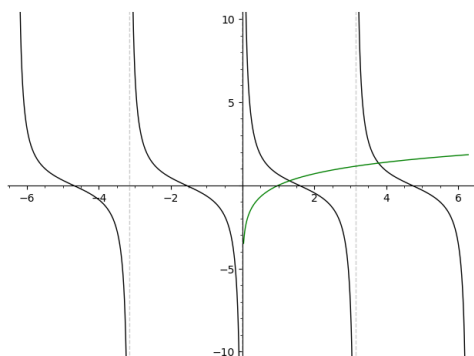
(f)  $\lim_{x \rightarrow 0} \left( \frac{\sin(x)}{x} \right)^{\frac{1}{x^2}} = \frac{1}{\sqrt[6]{e}}$ , (type  $1^\infty$ );

(g)  $\lim_{x \rightarrow 1^-} \left( \cos\left(\frac{\pi x}{2}\right) \right)^{\ln x} = 1$ , (type  $0^0$ ).

**Solution.** (a) The conclusion for the type is immediate. Set  $f(x) = \frac{\sin(2x) - 2 \sin(x)}{2e^x - x^2 - 2x - 2}$ . Applying repeatedly the l'Hopital's rule we obtain (observe that at any step the limit in the r.h.s exists)

$$\begin{aligned} \lim_{x \rightarrow 0} f(x) &\stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0} \frac{(\sin(2x) - 2 \sin(x))'}{(2e^x - x^2 - 2x - 2)'} \\ &= \lim_{x \rightarrow 0} \frac{2 \cos(2x) - 2 \cos(x)}{2e^x - 2x - 2} \\ &\stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0} \frac{(2 \cos(2x) - 2 \cos(x))'}{(2e^x - 2x - 2)'} \\ &= \lim_{x \rightarrow 0} \frac{-4 \sin(2x) + 2 \sin(x)}{2e^x - 2} \\ &\stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0} \frac{(-4 \sin(2x) + 2 \sin(x))'}{(2e^x - 2)'} \\ &= \lim_{x \rightarrow 0} \frac{-8 \cos(2x) + 2 \cos(x)}{2e^x} = -3. \end{aligned}$$

(b) Clearly,  $\lim_{x \rightarrow 0^+} \ln(x) = -\infty$  and  $\lim_{x \rightarrow 0^+} \cot(x) = +\infty$  (see also a sketch of the graphs of these functions below), hence the type of indeterminate form.



Otherwise, for a direct verification via Sage, one may type

```
g(x)=cot(x);k(x)=ln(x)
lim(g(x), x=0, dir="plus")
lim(ln(x), x=0, dir="plus")
```

This time applying the l'Hopital's rule, we will result to an indeterminate form of different type, which is again a quite common situation. In particular, recall that  $\cot(x) = \frac{1}{\tan(x)} = \frac{\cos(x)}{\sin(x)}$  and hence  $\cot'(x) = -1/\sin^2(x)$ . Setting  $f(x) = \ln(x)/\cot(x)$  and observing that at any step the limit in the

called *upper limit* of the sequence and denoted by

$$\limsup_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} \sup\{b_k; k \geq n\}.$$

The advantage is that the upper limit always exists (either finite if the sequence is bounded, or  $\pm\infty$  if the sequence is unbounded). Similarly,

$$\liminf_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} \inf\{b_k; k \geq n\}.$$

Therefore, we can reformulate the previous result (without having to change the proof) in a stronger form:

**Corollary.** Let  $S = \sum_{n=0}^{\infty} a_n$  be an infinite series of real or complex numbers.

(1) If

$$q = \limsup_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|,$$

then the series  $S$  converges absolutely for  $q < 1$  and does not converge absolutely for  $q > 1$ . For  $q = 1$ , it may or may not converge.

(2) If

$$q = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|},$$

the series converges absolutely for  $q < 1$  while it does not converge absolutely for  $q > 1$ . For  $q = 1$ , it may or may not converge.

In the literature, the ratio test is often called the *d'Alembert's criterion*, while the root test is called the *Cauchy's criterion*. There are many other useful tests, but we do not have space here for them.

**5.4.7. Alternating series.** The condition  $a_n \rightarrow 0$  is a necessary but not sufficient condition for the convergence of the series  $\sum_{n=0}^{\infty} a_n$ . However, there is the *Leibniz's criterion* of convergence for special types of series.

#### LEIBNIZ'S CRITERION FOR ALTERNATING SERIES

The series  $\sum_{n=0}^{\infty} (-1)^n a_n$ , where  $a_n$  is a non-increasing sequence of non-negative real numbers, is called an *alternating series*.

**Theorem.** An alternating series converges if and only if  $\lim_{n \rightarrow \infty} a_n = 0$ . Its value  $a = \sum_{n=0}^{\infty} (-1)^n a_n$  differs from the partial sum  $s_{2k}$  by at most  $a_{2k+1}$ .

**PROOF.** By the definition the partial sums  $s_k$  of an alternating series satisfy

$$s_{2(k+1)+1} = s_{2k+1} + a_{2k+2} - a_{2k+3} \geq s_{2k+1}$$

$$s_{2(k+1)} = s_{2k} - a_{2k+1} + a_{2k+2} \leq s_{2k}$$

$$s_{2k+1} - s_{2k} = -a_{2k+1} \rightarrow 0$$

$$s_2 \geq s_{2k} \geq s_{2k+1} \geq s_1.$$

Thus, the even partial sums are a non-decreasing sequence, while the odd ones are non-increasing. The last line reveals that the bounded sequence of the odd partial sums converges

r.h.s. exists, the following makes sense

$$\begin{aligned} \lim_{x \rightarrow 0^+} f(x) &\stackrel{\frac{+\infty}{+\infty}}{=} \lim_{x \rightarrow 0^+} \frac{\ln'(x)}{\cot'(x)} = \lim_{x \rightarrow 0^+} \frac{\frac{1}{x}}{-\frac{1}{\sin^2(x)}} \\ &= \lim_{x \rightarrow 0^+} \frac{-\sin^2(x)}{x} \stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0^+} \frac{-(\sin^2(x))'}{(x)'} \\ &= \lim_{x \rightarrow 0^+} (-2 \sin(x) \cos(x)) = 0. \end{aligned}$$

(c) For a verification of the type, notice that  $\lim_{x \rightarrow 1^+} \frac{x}{x-1} = +\infty$ , and  $\lim_{x \rightarrow 1^+} \frac{1}{\ln x} = +\infty$ . However, we see that  $f(x) := \left(\frac{x}{x-1} - \frac{1}{\ln x}\right) = \frac{x \ln(x) - x + 1}{(x-1) \ln(x)}$ , thus we obtain the indeterminate form  $0/0$ . One computes the following:

$$\begin{aligned} \lim_{x \rightarrow 1^+} f(x) &\stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 1^+} \frac{(x \ln(x) - x + 1)'}{((x-1) \ln(x))'} \\ &= \lim_{x \rightarrow 1^+} \frac{\ln(x) + \frac{x}{x} - 1}{\ln(x) + \frac{x-1}{x}} \\ &= \lim_{x \rightarrow 1^+} \frac{x \ln(x)}{x \ln(x) + x - 1} \\ &\stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 1^+} \frac{(x \ln(x))'}{(x \ln(x) + x - 1)'} = \frac{1}{2}. \end{aligned}$$

(d) Obviously, this case is of type  $0 \cdot (-\infty)$ . We transform the given expression into the type  $\frac{-\infty}{-\infty}$  by writing  $f(x) := \ln(x) \cdot \ln(x-1) = \frac{\ln(x-1)}{\frac{1}{\ln(x)}}$ . Then we see that

$$\begin{aligned} \lim_{x \rightarrow 1^+} f(x) &\stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 1^+} \frac{(\ln(x-1))'}{\left(\frac{1}{\ln(x)}\right)'} = \lim_{x \rightarrow 1^+} \frac{\frac{1}{x-1}}{-\frac{1}{x} \cdot \frac{1}{\ln^2(x)}} \\ &= \lim_{x \rightarrow 1^+} \frac{x \ln^2(x)}{1-x} \stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 1^+} \frac{(x \ln^2(x))'}{(1-x)'} \\ &= \lim_{x \rightarrow 1^+} \frac{\ln^2(x) + 2 \cdot \frac{x}{x} \cdot \ln(x)}{-1} = 0. \end{aligned}$$

(e) Recall from above that  $\lim_{x \rightarrow 0^+} \cot(x) = +\infty$ . Moreover  $\lim_{x \rightarrow 0^+} \frac{1}{\ln(x)} = \frac{1}{-\infty} = 0$ , hence indeed we have the indeterminate form  $(+\infty)^0$ . Moreover, we know that

$$\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^+} (\cot(x))^{\frac{1}{\ln(x)}} = e^{\lim_{x \rightarrow 0^+} \frac{\ln(\cot(x))}{\ln(x)}},$$

and hence it suffices to calculate the limit given in the argument of the exponential function. Setting  $g(x) = \frac{\ln(\cot(x))}{\ln(x)}$

to its supremum, while the even ones converge to the infimum. The previous line says they coincide, if and only if  $\lim_{n \rightarrow \infty} a_n = 0$  which proves the first claim.

At the same time the limit value  $a$  of the series is always at least  $s_{2k+1}$  and at most  $s_{2k}$ . Thus, the latter partial sums cannot differ by more than  $a_{2k+1}$  from the limit value.  $\square$

**Remark** (Riemann's rearrangement theorem). As obvious from the latter theorem, convergent alternating series are often not converging absolutely. This phenomenon is called *conditionally converging series*.

Unlike the independence on the order in which we sum up the terms of an absolutely convergent series (cf. (4) of Theorem 5.4.4), there is the famous Riemann theorem saying that conditionally convergent series can be brought to any finite or infinite value by appropriate rearrangement of the terms in the sum. We shall not go into the detailed proof here, but the idea is very simple: if a series  $S$  converges only conditionally, we may split it into the series of positive and negative terms,  $S_+$  and  $S_-$ , say ordered by absolute value, and both of them have to be divergent (otherwise, they would both converge absolutely and thus their difference would do as well). Now, prescribing the desired limit  $q$ , we shall add the positive terms until getting bigger than  $q$ , then adding the negative ones, until getting smaller than  $q$ , etc.

**5.4.8. Convergence rate.** The proofs of the tests derived in the previous two paragraphs allow also for straightforward estimates of the speed of the convergence. Indeed, both the tests for the absolute convergence are based on the comparison with the geometric series either for  $q = \limsup_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|$  or  $q = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$ , and  $0 < q < 1$ . In the estimate of the error of approximation of the limit  $s_\infty$  by the  $n$ -th partial sum  $s_n$

$$|s_\infty - s_n| < |a_N| \sum_{j=0}^{\infty} r^j = |a_N| r^{n-N} \frac{1}{1-r} = Cr^n$$

where  $N$  and  $q < r < 1$  are the two related choices from the proof of the test and  $C$  is the resulting constant not dependent of  $n$ . Thus the convergence rate is quite fast, in particular if  $r$  is much smaller than 1 (and we can get  $r$  as close to  $q$  as necessary).

On the other hand, the proof of the alternating series test shows that the convergence rate is at least as fast as the convergence of the terms  $a_n$ .

**5.4.9. Power series.** If we do not consider a sequence of numbers  $a_n$ , but rather a sequence of functions  $f_n(x)$  sharing the same domain  $A$ , we can use the definition of addition of series "point-wise", thereby obtaining the concept of the *series of functions*



$$S(x) = \sum_{n=0}^{\infty} f_n(x).$$



we obtain

$$\begin{aligned} \lim_{x \rightarrow 0^+} g(x) &\stackrel{+0/0}{=} \lim_{x \rightarrow 0^+} \frac{(\ln(\cot(x)))'}{\ln'(x)} \\ &= \lim_{x \rightarrow 0^+} \frac{\frac{1}{\cot(x)} \cdot (\cot(x))'}{\frac{1}{x}} \\ &= \lim_{x \rightarrow 0^+} x \cdot \left( \tan(x) \cdot \frac{-1}{\sin^2(x)} \right) \\ &= \lim_{x \rightarrow 0^+} \frac{-x}{\cos(x) \cdot \sin(x)} \\ &\stackrel{0/0}{=} \lim_{x \rightarrow 0^+} \frac{-1}{\cos^2(x) - \sin^2(x)} = -1. \end{aligned}$$

Thus  $\lim_{x \rightarrow 0^+} f(x) = e^{-1} = 1/e$ .

The cases (f) and (g) rely on the same trick as in the previous case. Thus,  $\lim_{x \rightarrow 0} \left(\frac{\sin(x)}{x}\right)^{\frac{1}{x^2}} = e^{\lim_{x \rightarrow 0} \frac{\ln(\frac{\sin(x)}{x})}{x^2}}$  and  $\lim_{x \rightarrow 1^-} \left(\cos\left(\frac{\pi x}{2}\right)\right)^{\ln(x)} = e^{\lim_{x \rightarrow 1^-} (\ln(x) \cdot \ln(\cos(\frac{\pi x}{2})))}$ , respectively, and now you should prove that the limits appearing as exponents in the right-hand-side equal to  $-1/6$  and  $0$ , respectively. Check yourself that this leads to the evaluations  $e^{-1/6} = 1/\sqrt[6]{e}$  and  $e^0 = 1$ , respectively.  $\square$

**5.C.34.** Show with an example that using the l'Hopital's rule for limits which are not indeterminate, can lead to wrong results.  $\circ$

### D. Infinite sums and power series



In this section we shall investigate whether we can add infinitely many real numbers, which leads us to consider what is known as an infinite series of real numbers. Usually we denote such series as  $\sum_{n=0}^{\infty} a_n$ .

Saying that an infinite series  $\sum_{n=0}^{\infty} a_n$  converges to a real number  $L$ , we mean that the sequence  $(S_n = \sum_{k=1}^n a_k)$  of partial sums converges to  $L$ , that is,

$$\lim_{n \rightarrow \infty} S_n = L.$$

If the sequence  $(S_n)$  does not converge, we usually say that the series diverges. When an infinite series  $\sum_{n=0}^{\infty} a_n$  is convergent, it is relatively easy to prove that  $a_n \rightarrow 0$  as  $n \rightarrow +\infty$ , see 5.4.5. Hence, if the sequence  $(a_n)$  does not converge to zero, then the given series is not convergent. However, there are many cases where  $a_n \rightarrow 0$  as  $n \rightarrow +\infty$ , but the induced series diverges, see our first example below (5.D.1).

In 5.4.5 further theorems are presented that help us to determine whether a series converges or diverges, known as “convergence criteria”. Many of these criteria revolve

If we consider the monomials  $f_n(x) = a_n x^n$ , we arrive at the “infinite polynomials”:

#### POWER SERIES

**Definition.** A power series is a series of functions given by the expression

$$S(x) = \sum_{n=0}^{\infty} a_n x^n$$

with coefficients  $a_n \in \mathbb{C}$ ,  $n = 0, 1, \dots$

$S(x)$  has the radius of convergence  $\rho \geq 0$  if and only if  $S(x)$  converges for every  $x$  satisfying  $|x| < \rho$  and does not converge for  $|x| > \rho$ .

**5.4.10. Properties of power series.** Although a significant part of the proof of the following theorem is postponed until the end of the following chapter, formulation of the basic properties of the power series can be considered now.

Actually, we should notice that our argument on the convergence of power series works exactly the same way for complex values of  $z$  and even now, the reader may enjoy a direct simple proof of all the claimed properties in the complex setting in 9.4.2 on page 873 in Chapter 9. Our real case may be viewed as a special case there and there is nothing included but some very straightforward and natural estimates (which could serve as a nice exercise right now).

Recall that the upper limit  $r = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$  equals the limit  $\lim_{n \rightarrow \infty} \sqrt[n]{|a_n|}$ , whenever this limit exists. Similarly with the ratio criterion and  $\limsup_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|$ .

#### CONVERGENCE AND DIFFERENTIATION

**Theorem.** Let  $S(x) = \sum_{n=0}^{\infty} a_n x^n$  be a power series and let

$$r = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}.$$

Then the radius of convergence of the series  $S$  is  $\rho = r^{-1}$ . Equivalently, we may compute

$$r = \limsup_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|.$$

The power series  $S(x)$  converges absolutely on the whole interval of convergence and is continuous on it (including the boundary points, supposing it is convergent there). Moreover, the derivative exists on this interval, and

$$S'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

**PROOF.** To verify the absolute convergence of the series, use the root test from theorem 5.4.5(3), for every value of  $x$ . Calculate (if the limit exists)

$$\lim_{n \rightarrow \infty} \sqrt[n]{|a_n x^n|} = r|x|.$$

around “absolute convergence”, a significant concept introduced in 5.4.4. Our goal next is to carefully analyze numerous examples to illustrate these theoretical statements. Additionally, we will demonstrate how Sage can be used for computational assistance. We start with a straightforward example, the so-called “harmonic series”, and progressively introduce more challenging examples and applications.

**5.D.1.** Explain why the following series are divergent:

$$S = \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}, \quad H = \sum_{n=1}^{\infty} \frac{1}{n}, \quad \mathcal{Y} = \sum_{n=1}^{\infty} \ln\left(\frac{n+1}{n}\right).$$

**Solution.** Starting with  $S$ , first of all observe that  $\frac{1}{\sqrt{n}} \rightarrow 0$  as  $n \rightarrow \infty$ . So indeed we cannot apply (1) of Theorem 5.4.5. However, considering the partial sums we see that

$$S_n = 1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}} > \frac{n}{\sqrt{n}} = \sqrt{n}$$

for all  $n \in \mathbb{N}^*$ . This means that the sequence  $(S_n)_{n \in \mathbb{N}^*}$  of partial sums of  $S$  is not bounded, and this implies that  $S$  diverges.

For  $H$  observe again that  $\frac{1}{n} \rightarrow 0$  as  $n \rightarrow \infty$ . This series is the *harmonic series*, discussed also in 5.B.19, and here we present a different proof verifying its divergence. The newer method is based on the partial sums of  $H$ , listed below:

$$\begin{aligned} H_2 &= 1 + \frac{1}{2} = H_1 + \frac{1}{2}, \\ H_4 &= H_2 + \frac{1}{3} + \frac{1}{4} \geq H_2 + \frac{1}{4} + \frac{1}{4} = H_2 + \frac{1}{2}, \\ H_8 &= H_4 + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \geq H_4 + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= H_4 + \frac{1}{2}, \\ \dots & \dots \\ H_{2^{n+1}} &= H_{2^n} + \frac{1}{2^n+1} + \frac{1}{2^n+2} + \cdots + \frac{1}{2^n+2^n} \\ &\geq H_{2^n} + \frac{2^n}{2^{n+1}} = H_{2^n} + \frac{1}{2}, \end{aligned}$$

and more general

$$H_{2^n} \geq 1 + \frac{n}{2}, \quad n \geq 1.$$

Thus, the sequence of partial sums is not bounded, and diverges to infinity. This implies that the harmonic series diverges as well. As we will see below, the harmonic series is a special case of a more general series, the so called “ $p$ -series”. For  $\mathcal{Y}$ , we have  $a_n := \ln\left(\frac{n+1}{n}\right) \rightarrow 0$  for  $n \rightarrow +\infty$ , as well. To verify this claim, use Sage via the cell

```
var("n"); lim(ln((n+1)/n), n=oo)
```

Either the series converges absolutely, or it does not converge if this limit is different from 1. It follows that it converges for  $|x| < \rho$  and diverges for  $|x| > \rho$ . If the limit does not exist, use the upper limit in the same way. The same argument based on the ratio test of convergence leads to the other formula for  $r$ .

The statements about continuity and the derivatives are proved later in a more general context, see 6.3.7–6.3.9, or check the straightforward proof in the complex setting in 9.4.2.  $\square$

**5.4.11. Remarks.** If the coefficients of the series increase rapidly enough, (for example  $a_n = n^n$ ), then  $r = \infty$ . Then the radius of convergence is zero, and the series converges only at  $x = 0$ .



Here are some examples of convergence of power series (including the boundary points of the corresponding interval): Consider

$$S(x) = \sum_{n=0}^{\infty} x^n, \quad T(x) = \sum_{n=1}^{\infty} \frac{1}{n} x^n.$$

The former example is the *geometric series*, which was already discussed. Its sum is, for every  $x$ ,  $|x| < 1$ ,

$$S(x) = \frac{1}{1-x},$$

while  $|x| > 1$  guarantees that the series diverges. For  $x = 1$ , we obtain the series  $1 + 1 + 1 + \dots$ , which is divergent. For  $x = -1$ , the series is  $1 - 1 + 1 - \dots$ , whose partial sums do not have a limit. The series oscillates.

Theorem 5.4.5(2) shows that the radius of convergence of the series  $T(x)$  is again 1 because

$$\lim_{n \rightarrow \infty} \left| \frac{\frac{1}{n+1} x^{n+1}}{\frac{1}{n} x^n} \right| = |x| \lim_{n \rightarrow \infty} \left| \frac{n}{n+1} \right| = |x|.$$

For  $x = 1$ , the series  $1 + \frac{1}{2} + \frac{1}{3} + \dots$ , is divergent: By summing up the  $2^{k-1}$  adjacent terms  $1/2^{k-1}, \dots, 1/(2^k - 1)$  and replacing each of them by  $2^{-k}$  (thus they total up to  $1/2$ ), the partial sums are bounded from below by the sum of these  $1/2$ 's. Since the bound from below diverges to infinity, so does the original series.

On the other hand, the series  $T(-1) = -1 + \frac{1}{2} - \frac{1}{3} + \dots$  converges although of course, it cannot converge absolutely. Of course, this is true since we deal here with an alternating series.

Notice that the convergence of a power series is relatively fast near  $x = 0$ . It is slower near the boundary of the convergence interval.

**5.4.12. Trigonometric functions.** Another important observation is that a power series is a series of numbers for each fixed  $x$  and the individual terms make sense for complex numbers  $x \in \mathbb{C}$ . Thus the domain of convergence of a power series is always a disc in the complex plane  $\mathbb{C}$  centered at the origin.



However, we may rely on the properties of natural logarithm, as follows:

$$\begin{aligned} \sum_{n=1}^{\infty} a_n &= \lim_{n \rightarrow \infty} \left( \ln \frac{2}{1} + \ln \frac{3}{2} + \ln \frac{4}{3} + \dots + \ln \frac{n+1}{n} \right) \\ &= \lim_{n \rightarrow \infty} \ln \frac{2 \cdot 3 \cdot 4 \cdot \dots \cdot (n+1)}{1 \cdot 2 \cdot 3 \cdot \dots \cdot n} \\ &= \lim_{n \rightarrow \infty} \ln(n+1) = +\infty. \end{aligned}$$

Thus the series  $\mathcal{Y}$  also diverges to  $+\infty$ .  $\square$

*The infinite number series*

$$\zeta(p) = \sum_{n=1}^{\infty} \frac{1}{n^p} = 1 + \frac{1}{2^p} + \dots + \frac{1}{n^p} + \dots$$

is called the  $p$ -series, and it is known that converges for all  $p > 1$ . For  $0 < p \leq 1$ , the  $p$ -series diverges (observe that for  $p = 1$  coincides with the harmonic series discussed above). For  $p > 1$  the  $p$ -series  $\zeta(p)$  is a monotone decreasing function of  $p$ , called the “Riemann zeta function”.<sup>9</sup> Euler proved that

$$\sum_{n=1}^{\infty} \frac{1}{n^p} = \prod_s \left( 1 + \frac{1}{s^p} \right),$$

where  $s$  ranges over all prime numbers. This formula, which is now known as the Euler product formula, makes  $\zeta$  a very important tool in number theory. However, for most of the  $p$  with  $p > 1$ , the value  $\zeta(p)$  is still unknown, though there are some accurate approximations (as for  $p = 3$  or  $p = 5$ , etc). Next we will use Sage to evaluate  $\zeta(p)$  for some  $p$ .

**5.D.2.** Evaluate the zeta function  $\zeta(p)$  for  $p = 2, \dots, 9$  with the aid of Sage. Is this feasible for all the given values of  $p$ ?  $\circ$

*Infinite sums naturally appear in many practical problems. Perhaps the best known ancient example is the Achilles paradox noticed by Aristoteles, which we briefly highlight by the next exercise. Later we will discuss more sophisticated problems, starting for example with 5.D.10.*

**5.D.3. Achilles’ paradox.** Although Achilles runs much faster than the turtle on the beach, he needs some time  $t_n$  for shortening their distance to the half. Thus Achilles needs the infinite sum of these time intervals  $t_n$  to catch the turtle. Explain why this sum is still a finite number.

**Solution.** If the current distance was  $d$ , and the velocity of the turtle was 1 while the velocity of Achilles was  $\alpha$ , we could write  $d + t - \alpha t$  for their distance after time  $t$ . Thus, shortening it to the half means  $t = \frac{1}{2} \frac{d}{\alpha - 1}$ . Repeating this  $n$  times, we obtain the time intervals

$$t_n = \frac{d}{\alpha - 1} \cdot \frac{1}{2^n}.$$

<sup>9</sup>Named for F. B. Riemann (1826 - 1866).

More generally, we can write power functions centered at an arbitrary (complex) point  $x_0$ ,

$$S(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n,$$

which converge absolutely again on the disc of radius  $\rho$ ,  $\rho^{-1} = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$ , but this time centered at  $x_0$ .

Earlier we proved explicitly (by a simple application of the ratio test) that the exponential function series converges everywhere. Thus this defines a function for all complex numbers  $x$ .

Its values are the limits of values of (complex) polynomials with real coefficients and each polynomial is completely determined by finitely many of its values. In particular, the values of each series are completely determined on the complex domain by their values at the real input values  $x$ . Therefore, the complex exponential must also satisfy the usual formulae which we have already derived for the real values  $x$ . In particular, for all  $x, y \in \mathbb{C}$

$$e^{x+y} = e^x \cdot e^y,$$

which can be also easily checked directly by the formula for the product in theorem 5.4.4(3).

Substitute the values  $x = it$ , where  $i \in \mathbb{C}$  is the imaginary unit,  $t \in \mathbb{R}$  arbitrary. We get the complex valued function on  $\mathbb{R}$

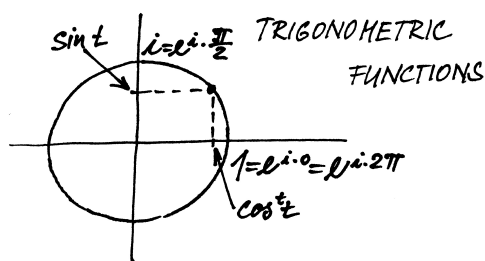
$$e^{it} = 1 + it - \frac{1}{2}t^2 - i\frac{1}{3!}t^3 + \frac{1}{4!}t^4 + i\frac{1}{5!}t^5 - \dots$$

The conjugate number to  $z = e^{it}$  is the number  $\bar{z} = e^{-it}$ . Hence

$$|z|^2 = z \cdot \bar{z} = e^{it} \cdot e^{-it} = e^0 = 1.$$

All the values  $z = e^{it}$  lie on the unit circle centered at the origin, in the complex plane.

The real and imaginary parts of the points lying on the unit circle are named as the trigonometric functions  $\cos \theta$  and  $\sin \theta$ , where  $\theta$  is the corresponding angle.



Differentiating the parametric description of the points of the circle  $t \mapsto e^{it}$ , gives the vectors of “velocities” which are easily computed. Differentiating the real and imaginary parts separately (knowing that the real power series can be differentiated term by term) gives :

$$\begin{aligned} (e^{it})' &= \left( 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 \dots \right)' + i \left( t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 \dots \right)' \\ &= -\left( t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 \dots \right) + i \left( 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 \dots \right) \end{aligned}$$

The procedure of summing all these  $t_n$  follows the rule of geometric series (see 5.B.16). Thus the result is

$$\sum_{n=0}^{\infty} \frac{1}{2^n} = \frac{1}{1 - \frac{1}{2}} = 2$$

and so Achilles needs the time  $\frac{d}{\alpha - 1}$  (here we sum from  $n = 1$ ). This perfectly matches the fact that their relative mutual velocity is  $\alpha - 1$  (explain why this assertion is true).  $\square$

*An infinite series, whether real or complex, that is absolutely convergent is also convergent. This is fundamentally due to the completeness of both the real numbers ( $\mathbb{R}$ ) and the complex numbers ( $\mathbb{C}$ ). Conversely, a series that is convergent but not absolutely convergent is termed “conditionally convergent”, see in the end of Section 5.4.7. In the following sections we will examine a range of examples that encompass all possible cases of convergence and divergence.*

**5.D.4.** Exam which of the following series is convergent or divergent:

$$\begin{aligned} T_1 &= \sum_{n=1}^{\infty} \frac{n^2 + 1}{2n^2 - 1}, & T_2 &= \sum_{n=1}^{\infty} \frac{2^n}{n}, \\ T_3 &= \sum_{n=1}^{\infty} \frac{2^n}{(n+2)!}, & T_4 &= \sum_{n=1}^{\infty} \frac{1}{n \cdot 2^{2025}}, \\ T_5 &= \sum_{n=1}^{\infty} \frac{n^7 + 7^n + 2}{(7^n + 2) \cdot n^7}, & T_6 &= \sum_{n=1}^{\infty} \frac{e^n}{n^2}, \\ T_7 &= \sum_{n=0}^{\infty} \frac{1}{(n+1) \cdot 3^n}, & T_8 &= \sum_{n=1}^{\infty} \frac{n^2 + 1}{n^3}, \\ T_9 &= \sum_{n=0}^{\infty} \frac{2^n + 3^n}{n!}, & T_{10} &= \sum_{n=1}^{\infty} 2^n \cdot \left(\frac{4}{5}\right)^{n^2}, \\ T_{11} &= \sum_{n=0}^{\infty} \frac{2 + \sin^3(n+1)}{4^n + n^2}, & T_{12} &= \sum_{n=1}^{\infty} \frac{n^n}{n^2 \cdot n!}. \end{aligned}$$

**Solution.** 1) We see that  $\lim_{n \rightarrow \infty} \frac{n^2+1}{2n^2-1} = \frac{1}{2} \neq 0$ , so by (1) in Theorem 5.4.5 the series  $T_1$  is divergent.

2) Similarly we see that  $\lim_{n \rightarrow \infty} \frac{2^n}{n} = \infty$ , hence  $T_2$  is divergent. Notice however that when the terms of the series are positive, absolute convergence is the same as convergence. Hence the same result follows by the ratio test, also known as the d’Alembert criterion (see (4) of Theorem 5.4.5). This is because

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n \rightarrow \infty} \left| \frac{\frac{2^{n+1}}{n+1}}{\frac{2^n}{n}} \right| = \lim_{n \rightarrow \infty} \frac{2(n+1)}{n} = 2 > 1.$$

One may like to verify the divergent of  $T_2$  (or  $T_1$ ) also in Sage. This can be done just by typing

```
var("n"); sum(2^n/n, n, 1, oo)
```

and what we should essentially keep from the answer that Sage returns is the very final output: **Sum is divergent.** Thus, the sum method in Sage can be used as a verification

which means  $(e^{it})' = i \cdot e^{it}$ . So the velocity vectors all have unit length. Hence the entire circle is parametrized if  $t$  is moved through the interval  $[0, 2\pi]$ , where  $2\pi$  stands for the length of the circle (a thorough definition of the length of a curve needs integral calculus, which we will develop in the next chapter). In particular, this procedure of parameterizing the circle can be used to define the *number*  $\pi$ , also called Archimedes’ constant or the Ludolphian number,<sup>15</sup> half the length of the unit circle in the Euclidean plane  $\mathbb{R}^2$ . It can be found by computing the first positive zero point of the imaginary part of  $e^{it}$ .

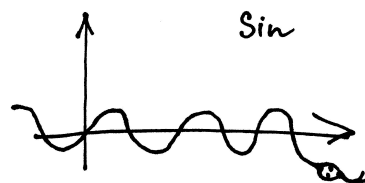
For example, use the 10th order approximation

$$\sin t \simeq t - \frac{1}{6}t^3 + \frac{1}{120}t^5 - \frac{1}{5040}t^7 + \frac{1}{362880}t^9.$$

Ask any computer algebra system to find the first positive root. The result is  $\pi \simeq 3.148690$ , for which only the first 2 decimal points are correct. But with the 20th order approximation this gets 3.141592 with 5 digits correct. The reason for the slow approximation is that  $\pi$  is relatively far from zero.

The explicit representation of trigonometric functions in terms of the power series is now apparent:

$$\begin{aligned} \cos t &= \operatorname{Re} e^{it} = 1 - \frac{1}{2}t^2 + \frac{1}{4!}t^4 - \frac{1}{6!}t^6 + \\ &\quad \dots + (-1)^k \frac{1}{(2k)!}t^{2k} + \dots \\ \sin t &= \operatorname{Im} e^{it} = t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \frac{1}{7!}t^7 + \\ &\quad \dots + (-1)^k \frac{1}{(2k+1)!}t^{2k+1} + \dots \end{aligned}$$



The following diagram illustrates the convergence of the series for the cosine function. It is the graph of the corresponding polynomial of degree 68. Drawing partial sums shows that the approximation near zero is very good. As the order increases, the approximation is better further away from the origin as well.

<sup>15</sup>This number describes the ratio of the circumference to the diameter of an (arbitrary) circle. It was known to the Babylonians and the Greeks in ancient times. The term Ludolphian number is derived from the name of German mathematician Ludolph van Ceulen of the 16th century, who produced 35 digits of the decimal expansion of the number, using the method of inscribed and circumscribed regular polygons, invented by Archimedes.

of divergence, as well.

3) We see that

$$\left| \frac{a_{n+1}}{a_n} \right| = \left| \frac{2^{n+1}(n+2)!}{2^n(n+3)!} \right| = \left| \frac{2}{n+3} \right| = \frac{2}{n+3},$$

which tends to  $0 < 1$  as  $n$  tends to infinity. By the ratio test we deduce that the series  $T_3$  is absolute convergent and hence also convergent.

4) Let  $\sum_{n=1}^{\infty} a_n$  be a non-convergent series. Taking some constant  $c \neq 0$  and considering the new series  $\sum_{n=1}^{\infty} (ca_n)$ , it is easy to see that this is also a non-convergent series. This is because by assumption the sequence of partial sums  $s_n = \sum_{k=1}^n a_k$  cannot converge, and hence also the sequence  $(S_n)$  of partial sums of the new series cannot converge (since  $S_n = \sum_{k=1}^n (ca_k) = c \cdot s_n$ ). We can apply this conclusion for  $T_4$ :

$$T_4 = \sum_{n=1}^{\infty} \frac{1}{n \cdot 2^{2025}} = \frac{1}{2^{2025}} \cdot \sum_{n=1}^{\infty} \frac{1}{n},$$

that is, the series is a multiple of the harmonic series. Thus, in combination with 5.D.1 we see that  $T_4$  is not convergent.

5) Obviously,

$$T_5 = \sum_{n=1}^{\infty} \frac{1}{n^7} + \sum_{n=1}^{\infty} \frac{1}{7^n + 2} = T_5^1 + T_5^2.$$

Hence it is sufficient to deduce about the convergence or divergence of the series  $T_5^1$  and  $T_5^2$ , as defined above. We see that  $T_5^2 = \sum_{n=1}^{\infty} \frac{1}{7^n + 2} < \sum_{n=1}^{\infty} \frac{1}{7^n}$  and the geometric series  $\sum_{n=1}^{\infty} \frac{1}{7^n}$  converges (see 5.B.16). Thus, the series  $T_5^2$  converges as well. The series  $T_5^1 = \sum_{n=1}^{\infty} \frac{1}{n^7}$  is also convergent (is a  $p$ -series with  $p = 7 > 1$ ). As a result, the relation  $T_5 = T_5^1 + T_5^2$  implies that  $T_5$  is a convergent series, (since it is the sum of two convergent series, see Corollary 5.4.3).

6) The series  $T_6$  is not convergent. Let us apply for example the d'Alembert criterion (ratio test):

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| &= \lim_{n \rightarrow \infty} \frac{e^{n+1} \cdot n^2}{e^n \cdot (n+1)^2} \\ &= e \cdot \lim_{n \rightarrow \infty} \left( \frac{n}{n+1} \right)^2 = e \cdot 1 = e > 1. \end{aligned}$$

7) The series  $T_7$  converges, because

$$T_7 = \sum_{n=0}^{\infty} \frac{1}{(n+1) \cdot 3^n} \leq \sum_{n=0}^{\infty} \left( \frac{1}{3} \right)^n = \frac{1}{1 - \frac{1}{3}} = \frac{3}{2} < +\infty.$$

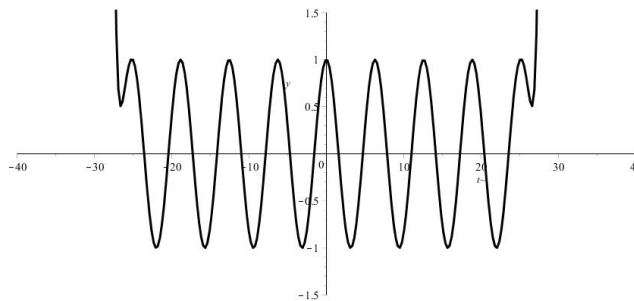
The result now follows by the comparison test (see Theorem 5.4.5).

8) The series  $T_8$  consists only of non-negative terms, as  $T_7$  for example, but it diverges (necessarily to  $+\infty$ ), as a direct computation shows:

$$T_8 = \sum_{n=1}^{\infty} \frac{n^2 + 1}{n^3} \geq \sum_{n=1}^{\infty} \frac{n^2}{n^3} = \sum_{n=1}^{\infty} \frac{1}{n} = +\infty.$$

9) Observe that  $T_9$  is again the sum of two series:

$$T_9 = \sum_{n=0}^{\infty} \frac{2^n}{n!} + \sum_{n=0}^{\infty} \frac{3^n}{n!} = T_9^1 + T_9^2$$



The well-known formula

$$e^{it} e^{-it} = \sin^2 t + \cos^2 t = 1$$

is immediate. From the derivative  $(e^{it})' = i e^{it}$  it follows that

$$(\sin t)' = \cos t, \quad (\cos t)' = -\sin t$$

by considering real and imaginary parts. Let  $t_0$  denote the smallest positive number for which  $e^{-it_0} = -e^{it_0}$ . The value  $t_0$  is the first positive zero point of the function  $\cos t$ . According to the definition of  $\pi$ ,  $t_0 = \frac{1}{2}\pi$ . If approximating the function  $\cos$  by the 10th order approximation, then twice the first positive root yields  $\pi \sim 3.1415917$  with 5 digits correct, while the 20th order approximation provides 10 correct digits after the decimal point.

Squaring yields  $e^{i2t_0} = e^{-i2t_0} = (e^{-it_0})^2$ . So  $\pi$  is a zero point of the function  $\sin t$ . Of course, for any  $t$ ,

$$e^{i(4kt_0+t)} = (e^{it_0})^{4k} \cdot e^{it} = 1 \cdot e^{it}.$$

Therefore, both trigonometric functions  $\sin$  and  $\cos$  are *periodic*, with period  $2\pi$ . This is their prime period.

Now the usual formulae connecting the trigonometric functions are easily derived. For illustration, we introduce some of them. First, the definition says that

$$\begin{aligned} (1) \quad \cos t &= \frac{1}{2}(e^{it} + e^{-it}) \\ (2) \quad \sin t &= \frac{1}{2i}(e^{it} - e^{-it}). \end{aligned}$$

Thus the product of these functions can be expressed as

$$\begin{aligned} \sin t \cos t &= \frac{1}{4i}(e^{it} - e^{-it})(e^{it} + e^{-it}) \\ &= \frac{1}{4i}(e^{i2t} - e^{-i2t}) = \frac{1}{2} \sin 2t. \end{aligned}$$

Further, by utilizing our knowledge of derivatives:

$$\cos 2t = \left( \frac{1}{2} \sin 2t \right)' = (\sin t \cos t)' = \cos^2 t - \sin^2 t.$$

The properties of other trigonometric functions

$$\tan t = \frac{\sin t}{\cos t}, \quad \cot t = (\tan t)^{-1}$$

can easily be derived from their definitions and the formulae for derivatives. The graphs of the functions sine, cosine, tangent, and cotangent are displayed on the diagrams (they are distinguished as the solid and dashed lines, respectively):

and hence it is sufficient to deduce about the convergence of  $T_9^1$  and  $T_9^2$ . By applying the ratio test to  $T_9^1$ , we see that

$$\lim_{n \rightarrow \infty} \left| \frac{2^{n+1}}{(n+1)!} \cdot \frac{n!}{2^n} \right| = \lim_{n \rightarrow \infty} \frac{2}{n+1} = 0 < 1$$

so  $T_9^1$  converges. Similarly for  $T_9^2$ , and hence  $T_9$  converges.

10) We can apply the root test, see (5) in Theorem 5.4.5:

$$\begin{aligned} \lim_{n \rightarrow \infty} \sqrt[n]{\left| 2^n \cdot \left(\frac{4}{5}\right)^{n^2} \right|} &= \lim_{n \rightarrow \infty} \sqrt[n]{2^n \cdot \left(\frac{4}{5}\right)^{n^2}} \\ &= \lim_{n \rightarrow \infty} \left( 2 \cdot \left(\frac{4}{5}\right)^n \right) = 0 < 1. \end{aligned}$$

Thus  $T_{10}$  is a convergent series. In Sage you can compute this limit as usual:

```
var("n"); lim(2*(4/5)**n, n=oo)
```

11) For the series  $T_{11}$  we see that

$$0 \leq \frac{2 + \sin^3(n+1)}{4^n + n^2} < \frac{3}{4^n}.$$

However,  $\sum_{n=1}^{\infty} \frac{3}{4^n} = 3 \sum_{n=1}^{\infty} \frac{1}{4^n}$  and the latter is a convergent geometric series. Thus the series  $T_{11}$  converges as well.

12) For the series  $T_{12}$  an application of the ratio test gives

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} &= \lim_{n \rightarrow \infty} \left( \frac{(n+1)^{n+1}}{(n+1)^2 \cdot (n+1)!} \cdot \frac{n^2 \cdot n!}{n^n} \right) \\ &= \lim_{n \rightarrow \infty} \frac{n^2}{(n+1)^2} \cdot \lim_{n \rightarrow \infty} \frac{(n+1)^n}{n^n} \\ &= \lim_{n \rightarrow \infty} \frac{n^2}{n^2} \cdot \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{n} \right)^n = 1 \cdot e > 1. \end{aligned}$$

Thus,  $T_{12}$  is not converge and in particular it diverges to  $+\infty$ .  $\square$

Next, we will explore how to program in Sage to utilize it as a tool for analyzing the convergence or divergence of infinite series. We introduce a syntax using the command `def` to construct a routine tailored for this purpose. The essence of this routine revolves around implementing the ratio test. Later on, we will demonstrate how this technique can be adapted to apply the root test (see 5.E.136), as well as other convergence criteria.



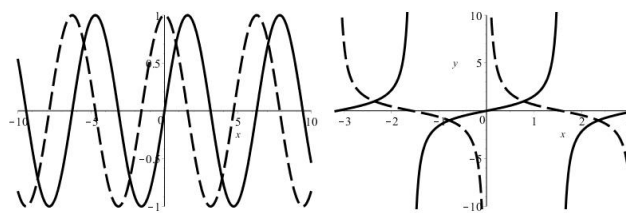
**5.D.5.** Let  $S = \sum_{n=0}^{\infty} f_n$  be an infinite series for which the limit  $\lim_{n \rightarrow \infty} |f_{n+1}/f_n|$  exists, and equals to  $q$ . Present in Sage a syntax appropriate to deduce the absolute convergence/divergence of  $S$  based on the ratio test. Then apply your program with aim to:

(a) Prove that the following infinite series converge:

$$S_1 = \sum_{n=1}^{\infty} \frac{2^n \cdot (n+1)^3}{3^n}, \quad S_2 = \sum_{n=1}^{\infty} \frac{6^n}{n!}.$$

Next provide a formal proof.

(b) Deduce, if possible, the convergence/divergence of  $S_3 =$



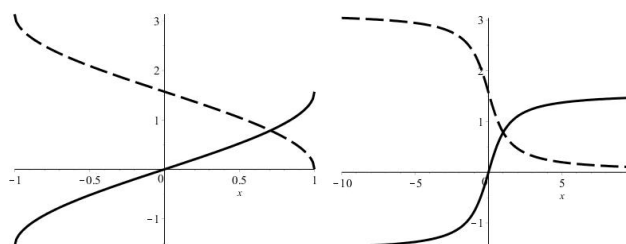
*Cyclometric functions* are the functions inverse to trigonometric functions. Since the trigonometric functions all have period  $2\pi$ , their inverses can be defined only inside one period, and further, only on the part where the given function is either increasing or decreasing. Two inverse trigonometric functions are

$$\arcsin = \sin^{-1}$$

with domain  $[-1, 1]$  and range  $[-\pi/2, \pi/2]$  and

$$\arccos = \cos^{-1}$$

with domain  $[-1, 1]$  and range  $[0, \pi]$ . See the left-hand illustration..



The remaining functions are (displayed in the diagram on the right)

$$\arctan = \tan^{-1}$$

with domain  $\mathbb{R}$  and range  $(-\pi/2, \pi/2)$ , and finally

$$\operatorname{arccot} = \cot^{-1}$$

with domain  $\mathbb{R}$  and range  $(0, \pi)$ .

The *hyperbolic functions* are also of big importance. Two basic ones are

$$\sinh x = \frac{1}{2}(e^x - e^{-x}), \quad \cosh x = \frac{1}{2}(e^x + e^{-x}).$$

The name indicates that they should have something in common with a hyperbola. From the definition,

$$(\cosh x)^2 - (\sinh x)^2 = 2 \frac{1}{4} (2e^x e^{-x}) = 1.$$

The points  $[\cosh t, \sinh t] \in \mathbb{R}^2$  parametrically describe a hyperbola in the plane. For hyperbolic functions, one can easily derive identities similar to the ones for trigonometric functions. By substituting into (1) and (2), one can obtain for example

$$\cosh x = \cos(ix), \quad i \sinh x = \sin(ix).$$

$\sum_{n=1}^{\infty} \frac{n^n}{n^2 \cdot n!}$  and  $S_4 = \sum_{n=1}^{\infty} \frac{n^2}{n+1}$  (notice  $S_3$  is the last series in 5.D.4).

**Solution.** As we said, we will use the command `def` to define a routine, which we will call `ratiotest`. First it is useful to introduce  $n$  as a symbolic variable. It is also convenient to introduce the subset  $(-1, 1) \subset \mathbb{R}$ . This is because of the ratio test (see Theorem 5.4.5), which states that  $S$  converges absolutely for  $|q| < 1$ , does not converge absolutely for  $|q| > 1$ , while for  $q = 1$  the series may or may not converge.

The aim now is to encode the above criterion in a Sage routine. For this we may type

```
var("n")
s=RealSet(-1, 1)
def ratiotest(f):
    q=lim(abs(f(n+1)/f(n)), n=oo)
    if q in s:
        return "converges absolutely"
    elif q==1:
        return "no conclusion"
    else:
        return "does not converge absolutely"
    return
```

Our routine is now ready to be tested.<sup>10</sup>

(a) First we will prove formally that  $S_1, S_2$  are both convergent series. Notice that for  $S_1, S_2$  the defining terms are positive, and hence

$$q = \lim_{n \rightarrow \infty} \left| \frac{f_{n+1}}{f_n} \right| = \lim_{n \rightarrow \infty} \frac{f_{n+1}}{f_n}.$$

So, for  $S_1$  we compute

$$\begin{aligned} q &= \lim_{n \rightarrow \infty} \frac{2^{n+1} \cdot (n+2)^3 \cdot 3^n}{3^{n+1} \cdot 2^n \cdot (n+1)^3} \\ &= \lim_{n \rightarrow \infty} \frac{2(n+2)^3}{3(n+1)^3} = \lim_{n \rightarrow \infty} \frac{2n^3}{3n^3} = \frac{2}{3} < 1. \end{aligned}$$

Or you can directly use Sage to compute this limit:

```
var("n")
f(n)=(2^(n+1)*(n+2)^3*3^n)/(3^(n+1)*(n+1)^3*2^n)
lim(f(n), n=oo)
```

Similarly, for  $S_2$ :

$$q = \lim_{n \rightarrow \infty} \left( \frac{6^{n+1}}{(n+1)!} \cdot \frac{n!}{6^n} \right) = \lim_{n \rightarrow \infty} \frac{6}{n+1} = 0 < 1,$$

where again one can compute the limit directly in Sage:

```
var("n")
g(n)=(factorial(n)*6^(n+1))/(factorial(n+1)*6^n)
lim(g(n), n=oo)
```

Notice here the command `factorial(n)` corresponds to  $n!$ . Since for both cases we obtained  $q < 1$ , our claim follows by the ratio test.

To check our routine using  $S_1, S_2$ , we type

<sup>10</sup>In general, detailed “return inputs” are more than welcome. However, here we kept them “short” to save some space.

**5.4.13. Notes.** (1) If a power series  $S(x)$  is expressed with the variable  $x$  moved by a constant offset  $x_0$ , we arrive at the function  $T(x) = S(x - x_0)$ . If  $\rho$  is the radius of convergence of  $S$ , then  $T$  will be well-defined on the interval  $(x_0 - \rho, x_0 + \rho)$ . We say that  $T$  is a *power series centered at  $x_0$* .



The power series can be defined in the following way:

$$S(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n,$$

where  $x_0$  is an arbitrary fixed real number. All of the previous reasonings are still valid. It is only necessary to be aware of the fact that they relate to the point  $x_0$ . Especially, such a power series converges on the interval  $(x_0 - \rho, x_0 + \rho)$ , where  $\rho$  is its radius of convergence.

Further, if a power series  $y = T(x)$  has its values in an interval where a power series  $S(y)$  is well-defined, then the values of the function  $S \circ T$  are also described by a power series which can be obtained by formal substitution of  $y = T(x)$  for  $y$  into  $S(y)$ .

(2) As soon as a power series with a suitable center is available, the coefficients of the power series for inverse functions can be calculated. We do not introduce a list of formulae here. It is easily obtained in Maple, for instance, by the procedure “series”. For illustration, here are two examples:

Begin with

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \dots$$

Since  $e^0 = 1$ , we search for a power series centered at  $x = 1$  for the inverse function  $\ln x$ . So assume

$$\ln x = a_0 + a_1(x-1) + a_2(x-1)^2 + a_3(x-1)^3 + a_4(x-1)^4 + \dots$$

Apply the equality  $x = \ln(e^x)$ , regroup the coefficients by the powers of  $x$  and substitute. The result is:

$$\begin{aligned} x &= a_0 + a_1 \left( x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \dots \right) \\ &\quad + a_2 \left( x + \frac{1}{2}x^2 + \dots \right)^2 + a_3 \left( x + \frac{1}{2}x^2 + \dots \right)^3 + \dots \\ &= a_0 + a_1 x + \left( \frac{1}{2}a_1 + a_2 \right) x^2 + \left( \frac{1}{6}a_1 + a_2 + a_3 \right) x^3 \\ &\quad + \left( \frac{1}{24}a_1 + \left( \frac{1}{4} + \frac{2}{6} \right) a_2 + \frac{3}{2}a_3 + a_4 \right) x^4 + \dots \end{aligned}$$

Comparing the coefficients of the corresponding powers on both sides, gives

$$a_0 = 0, \quad a_1 = 1, \quad a_2 = -\frac{1}{2}, \quad a_3 = \frac{1}{3}, \quad a_4 = -\frac{1}{4}, \dots$$

This corresponds to the valid expression (to be verified later):

$$\ln x = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} (x-1)^n.$$

Similarly, we can begin with the series

$$\sin t = t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \frac{1}{7!}t^7 + \dots$$

```
f(n) = ((2^n) * (n+1)^3) / 3^n; ratiotest(f)
```

and

```
f(n) = (6^n / factorial(n)); ratiotest(f)
```

respectively. Our routine verifies our result, by printing out the answer “converges absolutely”.

On the other hand, in 5.D.4 we saw that  $S_3$  satisfies  $q = e > 1$ , in particular,  $S_3$  cannot converge. The same verifies our routine:

```
f(n) = n^n / (n^2 * factorial(n)); ratiotest(f)
```

with answer “does not converge absolutely”. For the final case notice that running the command

```
print(lim((n+1)^3 / (n^2 * (n+2)), n=oo))
```

we obtain 1, that is,  $\lim_{n \rightarrow \infty} \left| \frac{f(n+1)}{f(n)} \right| = 1$ , where  $f(n) = \frac{n^2}{n+1}$ . This means that the ratio test is inconclusive and we cannot determine if the series converges or diverges using this test. Our routine is also able to certify this claim, and we just need to type

```
f(n) = n^2 / (n+1)
ratiotest(f)
```

which prints out “no conclusion”. To answer however this case, one may add the syntax `sum(n^2/(n+1), n, 1, oo)`, which shows to us that  $S_4$  is divergent. Try to describe a formal proof of this conclusion.  $\square$

**5.D.6. Alternating series.** Determine whether or not the following series converge:

$$S_1 = \sum_{n=1}^{\infty} (-1)^n \frac{n^2 + 3n - 1}{(3n - 2)^2},$$

$$S_2 = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{3n^4 - 3n^3 + 9n - 1}{(5n^3 - 2) \cdot 4^n}.$$

**Solution.** We have  $\lim_{n \rightarrow \infty} \frac{n^2 + 3n - 1}{(3n - 2)^2} = \lim_{n \rightarrow \infty} \frac{n^2}{9n^2} = \frac{1}{9} \neq 0$ . Thus the limit

$$\lim_{n \rightarrow \infty} (-1)^n \frac{n^2 + 3n - 1}{(3n - 2)^2}$$

does not exist, and the series  $S_1$  does not converge.

For  $S_2$  an application of the ratio (or root) test shows that the polynomials in the numerator or in the denominator do not affect the value of the considered limit, and we may consider the series

$$\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{4^n}$$

For this, we see that  $\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \frac{1}{4} < 1$ . Thus this series (absolutely) converges and as a conclusion the series  $S_2$  is also (absolutely) convergent.  $\square$

**5.D.7.** Decide whether the series

$$S_1 = \sum_{n=1}^{\infty} \frac{\sin(n)}{n^2}, \quad S_2 = \sum_{n=1}^{\infty} \frac{\cos(\pi n)}{\sqrt[3]{n^2}}$$

and the (unknown so far) power series for its inverse centered at zero (since  $\sin 0 = 0$ )

$$\arcsin t = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + \dots$$

Substitution gives

$$\begin{aligned} t &= a_0 + a_1 \left( t - \frac{1}{3!} t^3 + \frac{1}{5!} t^5 + \dots \right) + \\ & a_2 \left( t - \frac{1}{3!} t^3 + \frac{1}{5!} t^5 + \dots \right)^2 + \dots \\ &= a_0 + a_1 t + a_2 t^2 + \left( -\frac{1}{6} a_1 + a_3 \right) t^3 + \\ & \left( -\frac{2}{6} a_2 + a_4 \right) t^4 + \left( \frac{1}{120} a_1 - \frac{3}{6} a_3 + a_5 \right) t^5 + \dots, \end{aligned}$$

hence

$$\arcsin t = t + \frac{1}{6} t^3 + \frac{3}{40} t^5 + \dots$$

(3) Notice that if it is assumed that the function  $e^x$  can be expressed as a power series centered at zero, and that power series can be differentiated term by term, then the difference equation for the coefficients  $a_n$  is easily obtained since  $(x^{n+1})' = (n+1)x^n$ . Therefore, from the condition that the exponential function has its derivative equal to its value at every point,

$$a_{n+1} = \frac{1}{n+1} a_n, \quad a_0 = 1$$

and hence it is clear that  $a_n = \frac{1}{n!}$ .



converges absolutely, converges conditionally, or does not converge at all.  $\circ$

*Computing the exact values of infinite series is generally a challenging task. Sometimes, we can compare these values with known results. However, in many cases, relying on computer algebra software is necessary. To illustrate this situation, let us consider some examples and use Sage to verify our calculations rigorously.*

**5.D.8.** Calculate the series given below and then use Sage to verify your answer.

$$S_1 = \sum_{n=1}^{\infty} \left( \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \right), \quad S_2 = \sum_{n=0}^{\infty} \frac{5}{3^n},$$

$$S_3 = \sum_{n=1}^{\infty} \left( \frac{3}{4^{2n-1}} + \frac{2}{4^{2n}} \right), \quad S_4 = \sum_{n=1}^{\infty} \frac{n}{3^n}.$$

**Solution.** For the first given series we see that

$$\begin{aligned} S_1 &= \lim_{n \rightarrow \infty} \left( \left( \frac{1}{\sqrt{1}} - \frac{1}{\sqrt{2}} \right) + \left( \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{3}} \right) + \cdots \right. \\ &\quad \left. \cdots + \left( \frac{1}{\sqrt{n}} - \frac{1}{\sqrt{n+1}} \right) \right) \\ &= \lim_{n \rightarrow \infty} \left( 1 + \left( -\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \right) + \cdots \right. \\ &\quad \left. \cdots + \left( -\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right) - \frac{1}{\sqrt{n+1}} \right) \\ &= \lim_{n \rightarrow \infty} \left( 1 - \frac{1}{\sqrt{n+1}} \right) = 1. \end{aligned}$$

As we know, a verification by Sage takes the form

```
n=var("n"); sum((1/sqrt(n)-1/sqrt(n+1)), n, 1, oo)
```

For  $S_2$  we see that this is a quintuple of the standard geometric series with the common ratio  $q = 1/3$ . Thus

$$S_2 = \sum_{n=0}^{\infty} \frac{5}{3^n} = 5 \sum_{n=0}^{\infty} \left( \frac{1}{3} \right)^n = 5 \cdot \frac{1}{1 - \frac{1}{3}} = \frac{15}{2}.$$

Or we can directly give in Sage the cell

```
n=var("n"); sum((5/3**n), n, 0, oo)
```

The third infinite series is a series of linear combinations which we can express as a sum of infinite series with factoring out the constants. This is a valid modification, since the obtained series are absolutely convergent. In particular,

$$\begin{aligned} S_3 &= \frac{3}{4} \sum_{n=1}^{\infty} \left( \frac{1}{4^{2n-2}} \right) + \frac{2}{16} \sum_{n=1}^{\infty} \left( \frac{1}{4^{2n-2}} \right) \\ &\stackrel{m:=n-1}{=} \left( \frac{3}{4} + \frac{2}{16} \right) \sum_{m=0}^{\infty} \frac{1}{4^{2m}} \\ &= \frac{14}{16} \sum_{m=0}^{\infty} \left( \frac{1}{16} \right)^m = \frac{14}{16} \cdot \frac{1}{1 - \frac{1}{16}} = \frac{14}{15}. \end{aligned}$$

In Sage after introducing  $n$  as a symbolic variable we can compute this sum as above:

```
sum((3/(4**(2*n-1))+2/(4**(2*n))), n, 1, oo)
```

Finally, for  $S_4$ , notice that from the relation of the partial sums, i.e.,  $s_n = \frac{1}{3} + \frac{2}{3^2} + \frac{3}{3^3} + \dots + \frac{n}{3^n}$  for  $n = 1, 2, \dots$ , we can claim that

$$\frac{s_n}{3} = \frac{1}{3^2} + \frac{2}{3^3} + \dots + \frac{n-1}{3^n} + \frac{n}{3^{n+1}}.$$

Therefore, for all  $n \in \mathbb{Z}_+$  we obtain the relation

$$s_n - \frac{s_n}{3} = \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \dots + \frac{1}{3^n} - \frac{n}{3^{n+1}}.$$

This, in combination with the relation  $\lim_{n \rightarrow \infty} \frac{n}{3^{n+1}} = 0$ , gives

$$\begin{aligned} S_4 &= \frac{3}{2} \lim_{n \rightarrow \infty} \left( s_n - \frac{s_n}{3} \right) = \frac{3}{2} \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{1}{3^k} \\ &= \frac{3}{2} \sum_{k=1}^{\infty} \left( \frac{1}{3} \right)^k \stackrel{(*)}{=} \frac{3}{2} \left( \frac{1}{1 - \frac{1}{3}} - 1 \right) = \frac{3}{4}. \end{aligned}$$

Notice that here our geometric series  $\sum_{k=1}^{\infty} \left( \frac{1}{3} \right)^k$  starts from  $k = 1$  and thus the given replacement in  $(*)$ . A verification in Sage occurs as usual, i.e.,

```
var("n"); sum((n/3**n), n, 1, oo)
```

□

**5.D.9.** Verify the inequality  $\sum_{n=1}^{\infty} \frac{1}{n^2} < \sum_{n=0}^{\infty} \frac{1}{2^n}$ . ○

*Further exercises concerning infinite series are presented in Section E. Let us now describe a beautiful application that highlights the use of infinite series (see 5.E.130 for another fascinated example).*

**5.D.10. Koch snowflake (1904).** Create a “snowflake” by the following procedure: At the beginning, consider an equilateral triangle with sides of length 1. With each of its three sides, do the following: Cut it into three equally long parts, then build another equilateral triangle above the middle part (this is pointing out from the original triangle), and remove the middle part. This transforms the original equilateral triangle into a six-pointed star. Repeating this step ad infinitum, one arrives to the desired snowflake. Prove that the created figure has infinite perimeter. Then determine its area.

**Solution.** The perimeter of the original triangle is equal to 3. In each step, the perimeter increases by one third since three parts of every line segment are replaced with four equally long ones. Thus the snowflake’s perimeter can be expressed as the limit

$$\lim_{n \rightarrow \infty} d_n, \quad d_n := 3 \left( \frac{4}{3} \right)^n,$$

and we see that  $\lim_{n \rightarrow \infty} d_n = +\infty$ . This can be quickly verified in Sage:

```
var("n"); lim(3*((4/3)**n), n=oo)
```

Now, the figure's area is apparently increasing during the construction. For its computation it is sufficient to catch the rise between two consecutive steps. The number of the figure's sides is four times higher at every step (the line segments are divided into thirds and one of them is doubled). Moreover, the new sides are three times shorter. Therefore, the figure's area grows exactly by the equilateral triangles glued to each side (so, there is the same number of them as of the sides). In the first iteration (when creating the six-pointed star from the original triangle), the area grows by the three equilateral triangles with sides of length  $1/3$  (one third of the original sides' length). Let us denote the area of the original equilateral triangle by  $S_0$ . If we realize that shortening an equilateral triangle's sides three times makes its area decrease nine times, we get  $S_0 + 3 \cdot \frac{S_0}{9}$  for the area of the six-pointed star. Similarly, in the next step we obtain the area of the figure as

$$S_0 + 3 \cdot \frac{S_0}{9} + 4 \cdot 3 \cdot \frac{S_0}{9^2}.$$

It is easy now to deduce that the area  $E$  of the resulting snowflake equals the limit

$$\begin{aligned} E &= \lim_{n \rightarrow \infty} \left( S_0 + 3 \frac{S_0}{9} + 4 \cdot 3 \frac{S_0}{9^2} + \dots + 4^n \cdot 3 \frac{S_0}{9^{n+1}} \right) \\ &= S_0 \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{3} + \frac{1}{3} \cdot \frac{4}{9} + \dots + \frac{1}{3} \cdot \left( \frac{4}{9} \right)^n \right) \\ &= S_0 \left( 1 + \frac{1}{3} \lim_{n \rightarrow \infty} \left( 1 + \frac{4}{9} + \dots + \left( \frac{4}{9} \right)^n \right) \right) \\ &= S_0 \left( 1 + \frac{1}{3} \lim_{n \rightarrow \infty} \sum_{k=0}^n \left( \frac{4}{9} \right)^k \right) \\ &= S_0 \left( 1 + \frac{1}{3} \sum_{k=0}^{\infty} \left( \frac{4}{9} \right)^k \right) = S_0 \left( 1 + \frac{1}{3} \cdot \frac{1}{1 - \frac{4}{9}} \right), \end{aligned}$$

that is,  $E = \frac{8}{5} S_0$ . Thus, the snowflake's area is  $8/5$  of the area of the original triangle, i.e.,

$$\frac{8}{5} S_0 = \frac{8}{5} \cdot \frac{\sqrt{3}}{4} = \frac{2\sqrt{3}}{5}.$$

We mention that this snowflake is an example of an infinitely long curve which encloses a finite area.  $\square$

*So far, we have explored the concept of assigning a value to a sum of infinitely many numbers. Now, we shift our focus to sums involving infinitely many functions. Specifically, we can consider such series for each argument  $x$  of the functions  $f_n(x)$ , particularly focusing on sums of polynomials. These are known as "power series", which always converge on an interval (or a disc in the complex plane, for complex polynomials). Moreover, the radius of convergence of these series can be readily determined from their coefficients, which is either a non-negative real number or  $\infty$ .*



Later, in Chapters 6 and 7, we will revisit other types of function series and explore additional concepts of convergence.

**5.D.11.** Consider the series  $S = \sum_{n=1}^{\infty} \frac{(x+4)^n}{n \cdot 5^n}$ . Determine

all  $x \in \mathbb{R}$  for which  $S$  is convergent.

**Solution.** Having the ratio test in mind, let us compute the limit  $\lim_{n \rightarrow \infty} \left| \frac{f_{n+1}(x)}{f_n(x)} \right|$ , that is,

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{\frac{(x+4)^{n+1}}{(n+1) \cdot 5^{n+1}}}{\frac{(x+4)^n}{n \cdot 5^n}} \right| &= \lim_{n \rightarrow \infty} \frac{n |x+4|}{5(n+1)} \\ &= |x+4| \lim_{n \rightarrow \infty} \frac{n}{5(n+1)} = \frac{|x+4|}{5}. \end{aligned}$$

In case you like confirm this conclusion by Sage, give the cell `var("n"); lim(n/(5*(n+1)), n = oo)`.

Now, according to the ratio test the series will converge for  $\frac{|x+4|}{5} < 1$ , that is,  $-9 < x < 1$  and diverges for  $\frac{|x+4|}{5} > 1$ , that is,  $x < -9$  or  $x > 1$ . For  $x = 1$  we see that  $S$  is the harmonic series, i.e.,  $S = \sum_{n=1}^{\infty} \frac{1}{n}$  which diverges by 5.D.1. For  $x = -9$  we get

$$S = \sum_{n=1}^{\infty} \frac{(-5)^n}{n \cdot 5^n} = \sum_{n=1}^{\infty} \frac{(-1)^n}{n} = -1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \dots$$

which is an alternating series, known as the *alternating harmonic series*. For this series we compute  $\lim_{n \rightarrow \infty} \frac{(-1)^n}{n} = 0$ , hence  $S$  is convergent (see Theorem 5.4.7). As an alternative, in Sage the command `sum((-1)^n/n, n, 1, oo)` provides an explicit evaluation of the alternating harmonic series; it equals to  $-\ln(2)$ . We conclude that the initial series is convergent for all  $x \in [-9, 1)$ .  $\square$

**5.D.12.** Determine the radius  $r$  of convergence of the following power series:

$$\begin{aligned} A(x) &= \sum_{n=1}^{\infty} \frac{2^n}{n} x^n, & B(x) &= \sum_{n=1}^{\infty} \frac{1}{(1+i)^n} x^n, \\ C(x) &= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n \cdot 8^n} x^n, & D(x) &= \sum_{n=1}^{\infty} (-4n)^n x^n, \\ E(x) &= \sum_{n=1}^{\infty} \left(1 + \frac{1}{n}\right)^{n^2} x^n. \end{aligned}$$

**Solution.** According to the discussion in 5.4.10, for  $A(x)$  we have

$$r = \frac{1}{\limsup_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|} = \frac{1}{2}.$$

Thus, the power series converges exactly for the real numbers  $x \in (-\frac{1}{2}, \frac{1}{2})$ . Moreover, the series diverges for  $x = \frac{1}{2}$  (since it becomes harmonic), and it converges for  $x = -\frac{1}{2}$  (since then it becomes an alternating harmonic series). However, to determine the convergence for any  $x$  lying in the complex

plane on the circle of radius  $\frac{1}{2}$ , it is a much harder question which goes beyond our lectures.

For  $B(x)$  we compute

$$r = \limsup_{n \rightarrow \infty} \left| \sqrt[n]{\frac{1}{(1+i)^n}} \right| = \limsup_{n \rightarrow \infty} \left| \frac{1}{1+i} \right| = \frac{\sqrt{2}}{2}.$$

For  $C(x)$  we see that

$$\lim_{n \rightarrow \infty} \sqrt[n]{|a_n|} = \lim_{n \rightarrow \infty} \frac{1}{\sqrt[n]{n} \cdot 8} = \frac{1}{8}.$$

Thus the radius is  $r = 8$ . For  $D(x)$  we get  $\lim_{n \rightarrow \infty} \sqrt[n]{|a_n|} = \lim_{n \rightarrow \infty} 4n = +\infty$ , so the radius is  $r = 0$ , while for  $E(x)$  we compute  $\lim_{n \rightarrow \infty} \sqrt[n]{|a_n|} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e$ , so  $r = 1/e$  in this case.  $\square$

**5.D.13.** Calculate the radius  $r$  of convergence of the power series

$$\sum_{n=1}^{\infty} e^{in} \frac{\sqrt[3]{n^3 + n} \cdot 3^n}{\pi^n \cdot \sqrt[3]{n^4 + 2n^3 + 1}} \cdot (x - 2)^n.$$

**Solution.** Here we will apply the following trick: The radius of convergence of any power series does not change if we move its center or alter its coefficients while keeping their absolute values. Therefore, let us determine the radius of convergence of the infinite series

$$\sum_{n=1}^{\infty} \frac{\sqrt[3]{n^3 + n} \cdot 3^n}{\pi^n \sqrt[3]{n^4 + 2n^3 + 1}} \cdot x^n.$$

Since  $\lim_{n \rightarrow \infty} \sqrt[n]{n^a} = \left(\lim_{n \rightarrow \infty} \sqrt[n]{n}\right)^a = 1$  for all  $a > 0$ , we can further move to the series

$$\sum_{n=1}^{\infty} \frac{3^n}{\pi^n} x^n.$$

with the same radius of convergence  $r = \pi/3$ .  $\square$

**5.D.14.** Determine the power series centered at the origin which, determines the function

$$\frac{1}{x^2 - x - 12}, \quad x \in (-3, 3).$$

**Solution.** A quick method relies on the known procedure of partial fraction decomposition, which gives

$$\frac{1}{x^2 - x - 12} = \frac{1}{(x - 4)(x + 3)} = \frac{1}{7} \left( \frac{1}{x - 4} - \frac{1}{x + 3} \right).$$

This expression can be obtained also in Sage, via the `partial_fraction` function, as follows:

```
f(x)=1/(x^2-x-12)
show(f.partial_fraction())
```

Using now appropriately the known formula of geometric series, we see that

$$\begin{aligned}\frac{1}{x-4} &= -\frac{1}{4} \left( 1 + \frac{x}{4} + \frac{x^2}{4^2} + \cdots + \frac{x^n}{4^n} + \cdots \right), \\ \frac{1}{x+3} &= \frac{1}{3} \left( 1 - \frac{x}{3} + \frac{x^2}{3^2} + \cdots + \frac{(-x)^n}{3^n} + \cdots \right).\end{aligned}$$

Thus, altogether we obtain

$$\begin{aligned}\frac{1}{x^2 - x - 12} &= -\frac{1}{28} \sum_{n=0}^{\infty} \frac{x^n}{4^n} - \frac{1}{21} \sum_{n=0}^{\infty} \frac{(-x)^n}{3^n} \\ &= \sum_{n=0}^{\infty} \left( \frac{(-1)^{n+1}}{21 \cdot 3^n} - \frac{1}{28 \cdot 4^n} \right) x^n.\end{aligned} \quad \square$$

**5.D.15.** Determine the radius  $r$  of convergence of the power series  $\sum_{n=0}^{\infty} \frac{2^{2n} \cdot n!}{(2n)!} x^n$ .

**5.D.16.** Calculate the radius of convergence for  $\sum_{n=1}^{\infty} 2^{\sqrt{n}} x^n$ .

**5.D.17.** Find the domain of convergence of the power series  $\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{3^{\sqrt{n}}} x^n$ .

**5.D.18.** Determine for which  $x \in \mathbb{R}$  the power series  $\sum_{n=1}^{\infty} \frac{(-3)^n}{\sqrt{n^4 + 2n^3 + 111}} (x-2)^n$  converges.

**5.D.19.** Determine for which  $x \in \mathbb{R}$  the series

$$\sum_{i=1}^{\infty} \frac{1}{2^n \cdot n \cdot \ln(n)} x^{3n}$$

converges.

**5.D.20.** Determine all  $x \in \mathbb{R}$  for which the power series  $\sum_{i=1}^{\infty} \frac{x^{2n}}{n^2}$  is convergent.

*The polynomials are functions which we can easily evaluate and thus, the partial sums of power series offer a valuable method for approximating the values of functions. However, as we will see below, to be successful we need some good estimate on the convergence speed.*



*Especially for convergent alternating series*

$$S = \sum_{n=0}^{\infty} (-1)^n a_n = a_0 - a_1 + a_2 - a_3 + \cdots$$

with  $a_0 \geq a_1 \geq a_2 \geq \cdots \geq a_n \geq \cdots \geq 0$  and  $\sum_{n=0}^{\infty} (-1)^n a_n = L$ , where  $L$  is a finite real number, we can prove a very useful estimation error given by

$$|L - S_k| \leq a_{k+1}.$$

Here  $S_k = \sum_{n=0}^k (-1)^n a_n$  denotes the partial sum corresponding to  $S$ . Hence, summing only terms up to the  $k$ th term  $a_k$  and omitting all the remaining terms, the approximation

error will be at most large as  $a_{k+1}$ . Let us illustrate this situation via examples.

**5.D.21.** Approximate  $\cos(1)$  with an error strictly less than  $10^{-8}$ . Then use Sage to find the actual approximation error.

**Solution.** From Section 5.4.12 we know the following expression of  $\cos(x)$  in terms of power series:

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!},$$

for all  $x \in \mathbb{R}$ . Thus

$$\cos(1) = 1 - \frac{1}{2} + \frac{1}{4!} - \frac{1}{6!} + \frac{1}{8!} - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}.$$

According to the previous remark on alternating series, stopping at the term  $\frac{1}{(2n)!}$  we will obtain an error which equals at most  $\frac{1}{(2(n+1))!} = \frac{1}{(2n+2)!}$ . Thus, to find the number of terms which we need to approximate  $\cos(1)$  with an error strictly less than  $10^{-8}$ , it suffices to solve the inequality

$$\frac{1}{(2n+2)!} < 10^{-8}. \quad (*)$$

There are many ways to solve this inequality, and perhaps, the simplest one is by testing several values of  $n$ . As a solution, one gets that  $n = 5$  is the smallest positive integer satisfying (\*). For instance, in Sage a solution goes as follows:

```
n=1;
while (1/factorial(2*n+2) >= 10^-8 ):
    n = n+1
print(n)
```

and this answers 5. Here, with the second line we program Sage to try all values until (\*) is true, starting with  $n = 1$ . Hence, the desired solution is

$$\cos(1) \approx 1 - \frac{1}{2} + \frac{1}{4!} - \frac{1}{6!} + \frac{1}{8!} - \frac{1}{10!},$$

with an error less than  $10^{-8}$ . We may find the error of the approximation in Sage as follows:

```
a=1-1/2+1/factorial(4)-1/factorial(6) \
+1/factorial(8)-1/factorial(10)
print(N(cos(1))-N(a))
bool((N(cos(1))-N(a))<(1/factorial(12)))
```

Notice here we used a slash to break the first line. In this cell the print command gives the actual approximation error  $2.07625261428035e - 9$ , which approximately translates to  $2.077 \cdot 10^{-9}$ . The final command returns True, verifying that the error is not larger than the 6th term  $(-1)^n \frac{1}{(2n)!} \stackrel{n=5+1}{=} 1/12!$  (since above we proved that  $n = 5$ ), that is,  $1/12! \approx 2.088 \cdot 10^{-9}$ . For a successful treating of such small numbers one can use again Sage, e.g., by the cell

```
bool(2.077*(10**(-9))<2.088*(10**(-9)))
```

□

**5.D.22.** Approximate  $\sin(1)$  with an error strictly less than  $10^{-8}$ , and then verify your answer in Sage.

**Solution.** In Section 5.4.12 we learned the expression of  $\sin(x)$  in terms of power series:

$$\begin{aligned}\sin(x) &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}, \quad x \in \mathbb{R}.\end{aligned}$$

Thus

$$\sin(1) = 1 - \frac{1}{3!} + \frac{1}{5!} - \frac{1}{7!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}$$

and, as before, we are treating an alternating series. Stopping at the term  $\frac{(-1)^n}{(2n+1)!}$  we will obtain an error which equals at most  $\left| \frac{(-1)^{n+1}}{(2n+3)!} \right| = \frac{1}{(2n+3)!}$ . Hence, to locate the number of terms which we need to approximate  $\sin(1)$  with a cost strictly less than  $10^{-8}$ , it suffices to solve the inequality

$$\frac{1}{(2n+3)!} < 10^{-8},$$

and as a solution we get  $n = 5$  (e.g. by applying the same method as in 5.D.21). This means that

$$\sin(1) \approx 1 - \frac{1}{3!} + \frac{1}{5!} - \frac{1}{7!} + \frac{1}{9!} - \frac{1}{11!}$$

with an error less than  $10^{-8}$ . In Sage you may type

```
b=1-1/factorial(3)+1/factorial(5)-1/factorial(7) \
+1/factorial(9) -1/factorial(11)
print(abs(N(sin(1))-N(b)))
print(bool(abs(N(sin(1))-N(b))<1/10**8))
```

which returns the actual approximation error, given by  $1.59828483781155e-10 \approx 1.599 \cdot 10^{-10}$ , and verifies the statement.  $\square$



## E. Additional exercises for the whole chapter

As usual, we will now delve into additional material related to the concepts discussed thus far in Chapter 5. Many of the tasks described below rely on the theory of derivatives, making prior experience with derivatives particularly useful. In some instances, we may need to employ higher-order derivatives, which are formally introduced at the beginning of Chapter 6, see 6.1.1. However, these cases primarily pertain to higher-order derivatives of polynomials, a topic with which the reader is already familiar from our discussion in Section 5.1.6.



## A) MATERIAL ON POLYNOMIAL INTERPOLATION

**5.E.1.** Consider the function  $f(x) = \sin(x)$ . Given three points (nodes)  $x_0, x_1, x_2$ , write a routine in Sage which will return the Hermite interpolation polynomial  $P(x)$  corresponding to these nodes and the values  $y_i = \sin(x_i)$  and  $y'_i = \sin'(x_i) = \cos(x_i)$ , for  $i = 0, 1, 2$ , that is, adapted to the following table (for simplicity you may fix some triple  $(x_0, x_1, x_2)$ , e.g.,  $x_0 = -2\pi, x_1 = 0$  and  $x_2 = 2\pi$ )

$x_i$	$x_0$	$x_1$	$x_2$
$y_i$	$y_0 = \sin(x_0)$	$y_1 = \sin(x_1)$	$y_2 = \sin(x_2)$
$y'_i$	$y'_0 = \cos(x_0)$	$y'_1 = \cos(x_1)$	$y'_2 = \cos(x_2)$

Then, for a variety of different triples  $(x_0, x_1, x_2)$  centered to 0 (that is, with  $x_1 = 0$  and with  $x_0 = -x_2$ ), present the graphs of the corresponding polynomial  $P(x)$  and that of  $f(x)$ , for  $x \in [x_0, x_2]$ .

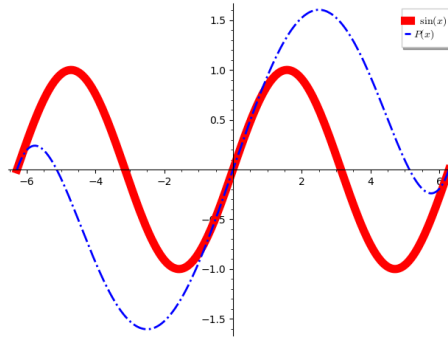
**Solution.** We present the code in Sage and attach some comments within. Since the Hermite interpolation method includes derivatives, below we will use commands as `derivative(f,x)`, which gives the derivative of a function  $f$  and `derivative(f,x,n)`, returning the  $n$ th derivative of  $f$ . So, let us fix the nodes  $x_0 = -2\pi, x_1 = 0, x_2 = 2\pi$ . Notice the elementary Lagrange polynomials  $\ell_0, \ell_1, \ell_2$  corresponding to these three nodes are all of degree 2. We have:

```
x0=-2*pi; x1=0; x2=2*pi; l(x)=(x-x0)*(x-x1)*(x-x2)
d1l(x)=derivative(l, x) #define first derivative of function l
d2l(x)=derivative(l(x), x, 2) #define second derivative of function l
l0(x)=((x-x1)*(x-x2))/((x0-x1)*(x0-x2)) #elementary Lagrange polynomial l_0
l1(x)=((x-x0)*(x-x2))/((x1-x0)*(x1-x2)) #elementary Lagrange polynomial l_1
l2(x)=((x-x0)*(x-x1))/((x2-x0)*(x2-x1)) #elementary Lagrange polynomial l_2
h01(x)=(1-(d2l(x0)/d1l(x0))*(x-x0))*(l0(x))^2 #1st type Hermite polyn. h^{(1)}_{0}
h11(x)=(1-(d2l(x1)/d1l(x1))*(x-x1))*(l1(x))^2 #1st type Hermite polyn. h^{(1)}_{1}
h21(x)=(1-(d2l(x2)/d1l(x2))*(x-x2))*(l2(x))^2 #1st type Hermite polyn. h^{(1)}_{2}
h02(x)=(x-x0)*(l0(x))^2 #2nd type Hermite polyn. h^{(2)}_{0}
h12(x)=(x-x1)*(l1(x))^2 #2nd type Hermite polyn. h^{(2)}_{1}
h22(x)=(x-x2)*(l2(x))^2 #2nd type Hermite polyn. h^{(2)}_{2}
y0=sin(x0); y1=sin(x1); y2=sin(x2) #introduce the values y_0, y_1, y_2
dy0=derivative(sin(x), x)(x=x0) #introduce the values y'_0
dy1=derivative(sin(x), x)(x=x1) #introduce the values y'_1
dy2=derivative(sin(x), x)(x=x2) #introduce the values y'_2
P(x)=y0*h01(x)+y1*h11(x)+y2*h21(x)+dy0*h02(x)+dy1*h12(x)+dy2*h22(x); show(P(x))
a=plot(sin(x), x, x0, x2, color="red", thickness=8, legend_label="$\sin(x)$")
b=plot(P(x), x, x0, x2, color="blue", thickness=2, linestyle="-.", legend_label="$P(x)$");
show(a+b)
```

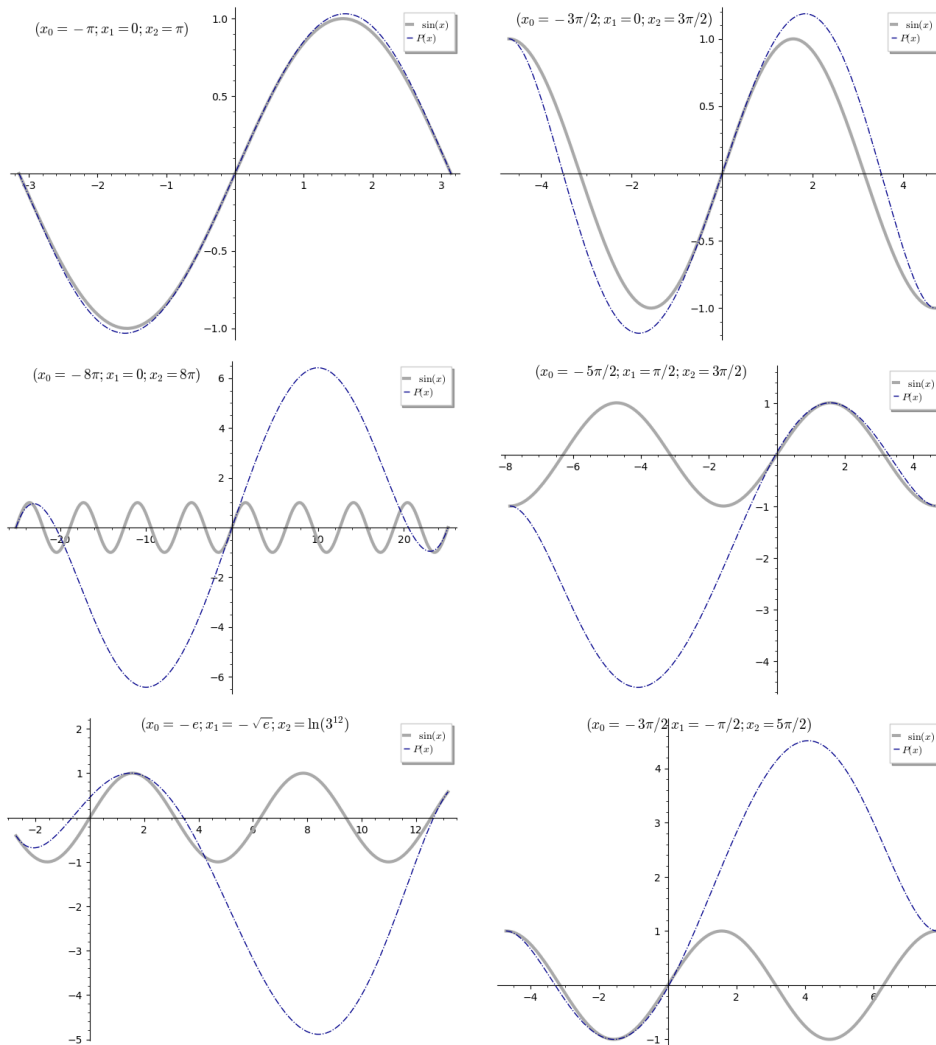
This returns the explicit form of the Hermite interpolation polynomial  $P(x)$ , given by

$$P(x) = \frac{(2\pi + x)^2(2\pi - x)^2x}{16\pi^4} - \frac{(2\pi + x)^2(2\pi - x)x^2}{64\pi^4} + \frac{(2\pi + x)(2\pi - x)^2x^2}{64\pi^4},$$

and the following figure



Above, the code is written in such way that one only changes the first line, that is, the initial values of  $x_0, x_1, x_2$ , (and so the triple  $(x_0, x_1, x_2)$ ) and results to a new Hermite interpolation polynomial. Let us present the graphs for the triples  $(-\pi, 0, \pi)$ ,  $(-3\pi/2, 0, 3\pi/2)$ , and  $(-8\pi, 0, 8\pi)$ , centered at 0, but also for other kind of triples, e.g.,  $(-5\pi/2, \pi/2, 3\pi/2)$ ,  $(-e, \sqrt{e}, \ln(3^{12}))$ , etc.



□

**5.E.2.** Determine the natural cubic spline that interpolates the absolute value function  $f(x) = |x|$ , for  $x \in [-1, 1]$ , selecting the points  $x_0 = -1, x_1 = 0, x_2 = 1$ . Then use the “spline” method described in 5.A.16 to verify your result.

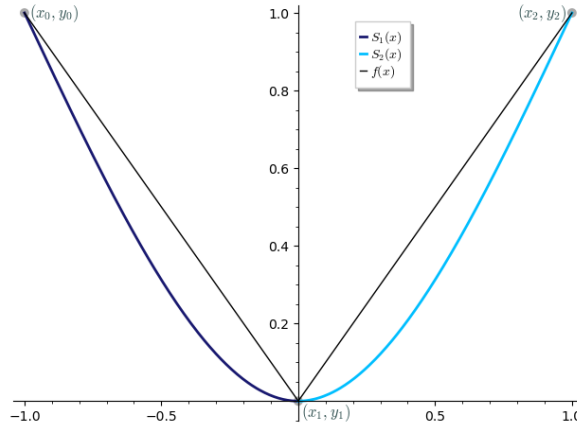
**Solution.** By assumption one has  $y_0 = f(x_0) = 1, y_1 = f(x_1) = 0$  and  $y_2 = f(x_2) = 1$ . For  $x \in [-1, 0]$  suppose that  $S(x) = S_1(x) = a_3x^3 + a_2x^2 + a_1x + a_0$  and for  $x \in [0, 1]$  let  $S(x) = S_2(x) = b_3x^3 + b_2x^2 + b_1x + b_0$ , with  $a_i, b_i \in \mathbb{R}$  for all

$i = 0, \dots, 3$ . The conditions that  $S(x)$  must satisfy are eight in total; First, we have  $S_1(x_0) = y_0, S_2(x_1) = y_1, S_1(x_1) = y_1$ , and  $S_2(x_2) = y_2$ , which equivalently are written as  $-a_3 + a_2 - a_1 + a_0 = 1, b_0 = 0 = a_0$  and  $b_3 + b_2 + b_1 + b_0 = 1$ , respectively. Moreover  $S_1'(x_1) = S_2'(x_1)$  and  $S_1''(x_1) = S_2''(x_1)$ , that is,  $a_1 = b_1$  and  $a_2 = b_2$ , respectively. Finally, since  $S$  should be natural, we need  $S_1''(x_0) = 0 = S_2''(x_2)$ , so we also have  $-6a_1 + 2a_2 = 0$  and  $6b_1 + 2b_2 = 0$ . The system of these eight linear equations has non-zero determinant and so unique solution. Using Sage we get  $a_0 = b_0 = 0, a_1 = b_1 = 0, a_2 = b_2 = 3/2, a_3 = 1/2$  and  $b_2 = -(1/2)$ , that is,

$$S(x) = \begin{cases} S_1(x) = (1/2)x^3 + (3/2)x^2, & \text{if } x \in [-1, 0], \\ S_2(x) = -(1/2)x^3 + (3/2)x^2, & \text{if } x \in [0, 1]. \end{cases}$$

Below we include the graphical verification obtained by a combination of the commands `spline` and `plot`, but present the figure that occurs by plotting  $S_1, S_2$  in Sage via the usual way (so one can compare alone in his/her editor).

```
f(x)=abs(x); pts=[(-1, f(-1)), (0, f(0)), (1, f(1))]; S=spline(pts)
A=plot(S, -1, 1, color="steelblue", thickness=2, legend_label="$S(x)$")
B=points(pts, size=50, color="darkgrey")
fpl=plot(f(x), x, -1, 1, color="black", legend_label="$f(x)$")
fpl.set_legend_options(loc=(0.6, 0.8))
tx0=text("$x_0, y_0$", (-0.89, 1), color="darkslategrey", fontsize="12")
tx1=text("$x_1, y_1$", (0.1, -0.03), color="darkslategrey", fontsize="12")
tx2=text("$x_2, y_2$", (0.89, 1), color="darkslategrey", fontsize="12")
show(A+B+fpl+tx0+tx1+tx2)
```



□

**5.E.3.** Without calculation, determine the Hermite interpolation polynomial if the following is given:

$$\begin{aligned} x_0 = 0, \quad x_1 = 2, \quad x_2 = 1, \\ y_0 = 0, \quad y_1 = 4, \quad y_2 = 1, \\ y'_0 = 0, \quad y'_1 = 4, \quad y'_2 = 2. \end{aligned}$$

○

**5.E.4.** Construct the natural cubic interpolation spline for the points  $x_0 = -3, x_1 = 0, x_2 = 3$ , and the values  $y_0 = -3, y_1 = 0, y_2 = 3$ .

○

**5.E.5.** Construct the complete cubic interpolation spline for the points  $x_0 = -3, x_1 = -2, x_2 = -1$  with values  $y_0 = 0, y_1 = 1, y_2 = 2$  and derivatives at the marginal points given by  $y'_0 = 1, y'_2 = 1$ .

○

**5.E.6.** Using Lagrange interpolation, approximate  $\cos^2(1)$ , based on the values of  $\cos^2(x)$  at the points  $\frac{\pi}{4}, \frac{\pi}{3}$ , and  $\frac{\pi}{2}$ .

○

**5.E.7.** Let  $P(x)$  be a polynomial with real non-negative coefficients. Assume that  $P(\frac{1}{x})P(x) \geq 1$  for  $x = 1$ . Show that the same inequality holds for every positive  $x$ .

**Solution.** Let  $P(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ . From the statement one has  $P(1)^2 \geq 1$ . Thus,

$$\begin{aligned} P(x)P\left(\frac{1}{x}\right) &= (a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0) \cdot (a_n x^{-n} + a_{n-1} x^{-(n-1)} + \dots + a_1 x^{-1} + a_0) \\ &= \sum_{i=0}^n a_i^2 + \sum_{i < j} a_i a_j (x^{j-i} + x^{i-j}) \geq \sum_{i=0}^n a_i^2 + 2 \sum_{i < j} a_i a_j = P(1)^2 \geq 1. \end{aligned}$$

Here we have used the well known inequality  $x + \frac{1}{x} \geq 2$ , which holds for any positive real number  $x$  (equivalent to  $(\sqrt{x} - \frac{1}{\sqrt{x}})^2 \geq 0$ ).  $\square$

**5.E.8.** Determine a polynomial  $P(x)$  of the least degree possible satisfying the conditions  $P(1) = 1$ ,  $P(2) = 28$ ,  $P(0) = 2$ ,  $P'(0) = 1$ ,  $P'(1) = 9$ .  $\circ$



Let us now explore material that focuses on the interpolation error of the Lagrange polynomial, which concerns the accuracy of interpolation. As we will soon demonstrate, Rolle's theorem (see 5.3.8), plays a crucial role in deriving this interpolation error. Later, in Chapter 6, we will observe that the formula for this error closely resembles that of the Taylor polynomial, see 6.1.3.

**5.E.9. Lagrange interpolation error.** Let  $x_0, \dots, x_n$  be  $(n + 1)$  distinct numbers (nodes) in an interval  $I = [a, b] \subset \mathbb{R}$ , and suppose that  $f(x)$  has  $(n + 1)$  derivatives on  $I$ , that is,  $f \in C^{n+1}[a, b]$ .<sup>11</sup> Assume also that  $f(x)$  is interpolated by the Lagrange polynomial  $P(x)$  passing through the  $(n + 1)$  given points. Prove that for each  $x \in I$  there exists a number  $\xi_x$  (generally unknown) in the open interval  $(a, b)$ , such that

$$f(x) - P(x) = \frac{\ell(x)}{(n + 1)!} f^{(n+1)}(\xi_x), \quad (*)$$

where  $f^{(n+1)}$  denotes the  $(n + 1)$ th derivative of  $f$  and by definition  $\ell(x) := (x - x_0)(x - x_1) \dots (x - x_n)$ . The difference  $f(x) - P(x)$  is called the *interpolation error*.

**Solution.** For a proof we will use repeatedly Rolle's theorem. For some  $t \in [a, b]$  distinct from all the numbers  $x_0, \dots, x_n$ , consider the function

$$g(x) := (f(x) - P(x)) - \frac{\ell(x)}{\ell(t)} \cdot (f(t) - P(t)).$$

Since by assumption  $f \in C^{n+1}[a, b]$ , it follows that  $g$  has also  $n + 1$  derivatives on  $I$ . Moreover, it is immediate that

$$g(t) = 0, \quad g(x_0) = 0, \quad \dots, \quad g(x_n) = 0,$$

which means that  $g$  has at least  $n + 2$  zeros in  $I$ . Recall now that Rolle's theorem states that between any two zeros of a differentiable function, exists at least one zero of its derivative. Thus  $g'$  must have at least  $n + 1$  zeros. We can again apply Rolle's theorem, and argue that the second derivative  $g''$  has at least  $n$  zeros if  $n \geq 1$ . Repeatedly we deduce that the third derivative  $g^{(3)}$  of  $g$  has  $n - 1$  zeros if  $n \geq 2$ , and finally  $g^{(n+1)}$  has at least one zero which we denote by  $\xi_t$ . Now,  $P$  is of degree  $n$  and its  $n + 1$  derivative vanishes,  $P^{(n+1)}(x) = 0$  for any  $x \in I$ . Moreover  $\ell(x)$  is polynomial of degree  $n + 1$ , and we see that  $\ell^{(n+1)}(x) = (n + 1)!$ . For instance for  $n = 2$ , that is,  $\ell(x) = (x - x_0)(x - x_1)(x - x_2)$  we have

$$\begin{aligned} \ell'(x) &= (x - x_1)(x - x_2) + (x - x_0)(x - x_1) + (x - x_0)(x - x_1), & \ell''(x) &= \\ &2(x - x_0) + 2(x - x_1) + 2(x - x_2), & \ell^{(3)}(x) &= 6 = (2 + 1)!. \end{aligned}$$

Also  $\ell(t)$  is a constant. Hence, differentiating  $(n + 1)$  times the function  $g(x)$  and evaluating the result at  $\xi_t$  yields the formula

$$0 = g^{(n+1)}(\xi_t) = f^{(n+1)}(\xi_t) - \frac{(n + 1)! (f(t) - P(t))}{\ell(t)}.$$

Upon solving for  $f(t)$  and replacing  $t$  by  $x$  and  $\xi_t$  by  $\xi_x$ , we get the result.  $\square$

Suppose that  $\hat{x}$  is a point at which we want to compute the error estimate. Then  $\xi = \xi_{\hat{x}}$  depends on that point (not in a continuous way in general), and is hardly known. Hence, having small control of  $f^{(n+1)}(\xi)$  one can claim that the formula (\*) is of limited utility. However, it can be used to obtain a bound on the interpolation error, as we explain below via examples.

**5.E.10.** Consider the function  $f(x) = x \ln(x)$  and the points  $x_0 = 8.3$ ,  $x_1 = 8.6$ . Compute the first-order Lagrange interpolation polynomial  $P$  corresponding to this data and next approximate  $f(8.4)$ . Moreover, estimate an upper bound of the absolute value  $|f(8.4) - P(8.4)|$  using the relation (\*), and compare the result with the real interpolation error.

**Solution.** Let us use Sage to compute  $f(x_0)$  and  $f(x_1)$ :

```
x0=8.3; x1=8.6; f(x)=x*ln(x); print([n(f(x0), digits=4), n(f(x1), digits=4)])
```

which returns [17.65, 18.51], that is,  $f(x_0) \approx 17.65$  and  $f(x_1) \approx 18.51$ , respectively. We have

$$\ell_0(x) = \frac{x - 8.6}{8.3 - 8.6} = -\frac{x - 8.6}{3/10} = -\frac{10}{3}(x - 8.6) = \frac{86 - 10x}{3}, \quad \ell_1(x) = \frac{x - 8.3}{8.6 - 8.3} = \frac{x - 8.3}{3/10} = \frac{10x - 83}{3}.$$

<sup>11</sup>Here we follow the notation of the paragraph 6.1.1.

Thus  $P(x) = f(x_0) \cdot \ell_0(x) + f(x_1) \cdot \ell_1(x) \approx 3.13412x + 8.44823$ , and we see that  $P(8.4) \approx 17.87834$ , which gives the required approximation of  $f(8.4)$ . The real value  $f(8.4)$  is approximately given by 17.87716, thus the absolute value of the interpolation error is  $|f(8.4) - P(8.4)| \approx 0.00119$ . These computations can be verified in Sage:

```
x0=8.3; x1=8.6; f(x)=x*ln(x); l0(x)=(x-x1)/(x0-x1); l1(x)=(x-x0)/(x1-x0)
P(x)=f(x0)*l0(x)+f(x1)*l1(x); abs(f(8.4)-P(8.4))
```

Or alternatively, one can proceed using the `lagrange.polynomial` command:

```
x0=8.3; x1=8.6; f(x)=x*ln(x); pt=[(x0, f(x0)), (x1, f(x1))]
R=PolynomialRing(QQ, "x"); p=R.lagrange_polynomial(pt); abs(f(8.4)-p(8.4))
```

Let us now estimate an upper bound of the error. By the formula (\*), and since  $\ell(x) = (x - x_0)(x - x_1) = (x - 8.3)(x - 8.6)$  and  $n = 1$ , we obtain

$$|f(8.4) - P(8.4)| = \left| \frac{f''(\xi)}{2!} (8.4 - 8.3)(8.4 - 8.6) \right| \quad (**)$$

where here  $\xi \equiv \xi_{8.4} \in (8.3, 8.6)$ . By the product rule we see that  $f'(x) = (x)' \cdot \ln(x) + x \cdot (\ln(x))' = \ln(x) + x \cdot \frac{1}{x} = \ln(x) + 1$ , so  $f''(x) = \frac{1}{x}$ . Consequently, we deduce that  $|f''(x)| \leq \frac{1}{8.3}$  for any  $x \in [8.3, 8.6]$  (and hence also for  $x = \xi$ ). Thus (\*\*) gives

$$|f(8.4) - P(8.4)| \leq \frac{1/8.3}{2} |8.4 - 8.3| \cdot |8.4 - 8.6| = \frac{0.1 \cdot 0.2}{16.6} \approx 0.001205.$$

Observe that the actual error  $|f(8.4) - P(8.4)| \approx 0.00119$  computed above, is a bit smaller than this upper bound, as it should be.  $\square$

**5.E.11.** Consider the exponential function  $f(x) = e^x$  and the points  $x_0 = 1, x_1 = 1.3, x_2 = 1.6, x_3 = 1.9, x_4 = 2$ . Suppose that  $P(x)$  is the Lagrange polynomial interpolating  $f$  along these nodes.

(a) What is the degree of the polynomial  $P$ ? Use Sage to determine  $P$ .

(b) Find an upper bound of the absolute value of the interpolation error when we approximate  $f(1.25)$  by  $P(1.25)$ .

**Solution.** (a) The Lagrange polynomial is at most of degree 4. In particular, via Sage and the cell below we see that it has degree exactly 4.

```
x0=1; x1=1.3; x2=1.6; x3=1.9; x4=2; f(x)=e**x
pt=[(x0, f(x0)), (x1, f(x1)), (x2, f(x2)), (x3, f(x3)), (x4, f(x4))]
R=PolynomialRing(RR, "x"); p=R.lagrange_polynomial(pt)
```

(b) Based on the relation (\*) we obtain

$$|f(x) - P(x)| = \left| \frac{f^{(5)}(\xi)}{(4+1)!} \right| |(x-1) \cdot (x-1.3) \cdot (x-1.6) \cdot (x-1.9) \cdot (x-2)|.$$

Since  $f(x) = e^x$ , we have  $f^{(5)}(x) = e^x$ , as well. Moreover we see that  $|f^{(5)}(x)| \leq e^2$  for any  $x \in [0, 2]$ , and hence  $|e^\xi| \leq e^2$ . Thus

$$\begin{aligned} |f(1.25) - P(1.25)| &= \left| \frac{e^\xi}{5!} \right| \cdot |1.25 - 1| \cdot |1.25 - 1.3| \cdot |1.25 - 1.6| \cdot |1.25 - 1.9| \cdot |1.25 - 2| \\ &\leq \frac{e^2}{120} (0.25) \cdot (0.05) \cdot (0.35) \cdot (0.65) \cdot (0.75) \approx \frac{7.38906}{120} \cdot 0.00214 \approx 0.000132. \end{aligned}$$

The absolute value of the actual error is approximately  $|f(1.25) - P(1.25)| \approx 0.0000810$ , and this is smaller of the upper bound given above. One can obtain the actual error in Sage by typing in the end of previous cell `print(abs(f(1.25) - p(1.25)))`.  $\square$

## B) MATERIAL ON REAL NUMBERS AND LIMIT PROCESSES

**5.E.12.** Determine the supremum and infimum of the set  $A = (-3, 0] \cup (1, \pi) \cup \{6\} \subset \mathbb{R}$ .

**Solution.** Clearly,  $\sup A = 6, \inf A = -3$ . One can verify this in Sage by applying the following the cell (notice however that this method applies only for subsets of  $\mathbb{R}$ )

```
A = RealSet.open_closed(-3, 0) + RealSet(1, pi) + RealSet.point(6)
print(A.inf()); print(A.sup())
```

$\square$

**5.E.13.** (a) Find  $\sup A$  and  $\inf A$  for

$$A = \left\{ a_n = \frac{n + (-1)^n}{n} : n \in \mathbb{N}^* \right\} \subset \mathbb{R}.$$

(b) Find a subset of  $\mathbb{R}$  which does not have an infimum in  $\mathbb{R}$ , but has a supremum.

(c) Find a subset of  $\mathbb{R}$  which does not have a supremum in  $\mathbb{R}$ , but has an infimum.

**Solution.** (a) Clearly

$$\begin{aligned} A &= \left\{ a_1 = 0, a_2 = \frac{3}{2}, a_3 = \frac{2}{3}, a_4 = \frac{5}{4}, \dots \right\} \\ &= \left\{ a_1 = 1 - \frac{1}{2 \cdot 1 - 1}, a_2 = 1 + \frac{1}{2 \cdot 1}, a_3 = 1 - \frac{1}{2 \cdot 2 - 1}, a_4 = 1 + \frac{1}{2 \cdot 2}, \dots \right\}. \end{aligned}$$

Therefore, the given sequence  $(a_n = \frac{n+(-1)^n}{n})_{n=1}^\infty$  splits into two subsequences with even and odd indices, that is,

$$(a_n)_{n=1}^\infty = (b_n)_{n=1}^\infty \cup (c_n)_{n=1}^\infty, \quad \text{with } b_n = a_{2n} = 1 + \frac{1}{2n}, \quad \text{and } c_n = a_{2n-1} = 1 - \frac{1}{2n-1},$$

respectively. Clearly  $\frac{3}{2} = b_1 \geq b_n > c_n \geq c_1 = 0$ . Thus,  $\sup A = \frac{3}{2}$  and  $\inf A = 0$ , respectively.

(b) An example is given by the set  $\mathbb{Z} \setminus \mathbb{N}$ .

(c) The set of natural numbers is such an example. □

**5.E.14.** Recall that when the supremum of a subset  $A \subset \mathbb{R}$  belongs to  $A$  is called “*maximum of A*”. Similarly is defined the “*minimum of A*”. Find, if existent, the minimum/maximum of the following sets:  $A = (-1, 2)$ ,  $B = (-1, 2]$ ,  $C = \{(-1)^n : n \in \mathbb{Z}_+\}$ , and  $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$ .

**Solution.** The set  $A = (-1, 2)$  has 2 as its supremum, but  $2 \notin A$ . So there is no maximum for  $A$ . Neither it has a minimum, since its infimum  $-1 \notin A$ . The set  $B = (-1, 2]$  has 2 as its supremum, which belongs to  $B$ . Hence 2 is the maximum of  $B$ . But it does not have a minimum. For  $C$  both the minimum and maximum elements exist, and are the numbers  $-1$  and  $1$ , respectively. Finally,  $\mathbb{R}_+$  is not bounded above, but does has an infimum which equals to 0. However,  $0 \notin \mathbb{R}_+$  hence in this case there is neither a minimum, nor a maximum. □

**5.E.15.** Find a subset  $X \subset \mathbb{R}$  such that  $\sup X \leq \inf X$ . ○

**5.E.16.** Find sets  $A, B, C \subseteq \mathbb{R}$  such that  $A \cap B = \emptyset$ ,  $A \cap C = \emptyset$ ,  $B \cap C = \emptyset$ , and  $\sup A = \inf B = \inf C = \sup C$ . ○

**5.E.17.** Based on the definition of a convergent sequence, show that  $a_n \rightarrow 1$ , as  $n \rightarrow \infty$ , where  $a_n = \frac{2^n - 1}{2^n}$ .

**Solution.** We have  $|a_n - 1| = \left| \frac{2^n - 1}{2^n} - 1 \right| = \left| \frac{1}{2^n} \right| = \frac{1}{2^n}$ . On the other hand, by induction we get  $n < 2^n$  for any natural number  $n$ , hence we also have  $\frac{1}{2^n} < \frac{1}{n}$  for all  $n \in \mathbb{Z}_+$ . Let  $\varepsilon > 0$ . Using the Archimedean property we can find some  $N$  with  $\frac{1}{N} < \varepsilon$ . Then, for  $n > N$  we get  $\frac{1}{n} < \frac{1}{N} < \varepsilon$  and hence  $|a_n - 1| = \frac{1}{2^n} < \frac{1}{n} < \frac{1}{N} < \varepsilon$ . This shows that  $a_n \rightarrow 1$  as  $n \rightarrow \infty$ . □

**5.E.18.** Use the binomial theorem to prove that  $\lim_{n \rightarrow +\infty} \sqrt[n]{\ln(n)} = 1$ .

**Solution.** For  $x > 0$  the function  $\ln(x)$  is strictly increasing, thus for  $n > 3$  it is true that  $\ln(n) > \ln(e) = 1$  and this implies that  $\sqrt[n]{\ln(n)} > 1$ . Consequently, we may write  $\sqrt[n]{\ln(n)} = 1 + a_n$  for some  $a_n > 0$  for all  $n$ . Taking the  $n$ th power of this relation we arrive to  $\ln(n) = (1 + a_n)^n$ , and here is where the binomial theorem applies:

$$\ln(n) = (1 + a_n)^n = 1 + na_n + \frac{n(n-1)}{2!}a_n^2 + \dots + a_n^n.$$

Thus  $\ln(n) > 1 + na_n$ , from where we deduce that  $a_n < \frac{\ln(n)}{n} - \frac{1}{n}$ . However,  $\frac{\ln(n)}{n} \rightarrow 0$  as  $n \rightarrow +\infty$ , thus  $\lim_{n \rightarrow +\infty} a_n = 0$ . Hence by the relation  $\sqrt[n]{\ln(n)} = 1 + a_n$  the result follows. □

**5.E.19.** Compute the limit  $\lim_{n \rightarrow \infty} (\sqrt{2} \cdot \sqrt[4]{2} \cdot \sqrt[8]{2} \dots \sqrt[2^n]{2})$ . ○

**5.E.20.** Evaluate the limit  $\lim_{n \rightarrow \infty} \left( \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \dots + \frac{1}{(n-1) \cdot n} \right)$ . ○

**5.E.21.** Compute the limit  $\lim_{n \rightarrow \infty} \left( \frac{1}{n^2} + \frac{2}{n^2} + \dots + \frac{n-2}{n^2} + \frac{n-1}{n^2} \right)$  ○

**5.E.22.** Evaluate the limit  $\lim_{n \rightarrow \infty} \left( \frac{1}{\sqrt{n^2+1}} + \frac{1}{\sqrt{n^2+2}} + \cdots + \frac{1}{\sqrt{n^2+n}} \right)$ . Also, use Sage to plot several terms of the given sequence to visually illustrate your answer.

**Solution.** In order to determine this limit, we can invoke the squeeze theorem. The bounds

$$\frac{1}{\sqrt{n^2+1}} + \cdots + \frac{1}{\sqrt{n^2+n}} \geq \frac{1}{\sqrt{n^2+n}} + \cdots + \frac{1}{\sqrt{n^2+n}} = \frac{n}{\sqrt{n^2+n}}$$

and

$$\frac{1}{\sqrt{n^2+1}} + \cdots + \frac{1}{\sqrt{n^2+n}} \leq \frac{1}{\sqrt{n^2+1}} + \cdots + \frac{1}{\sqrt{n^2+1}} = \frac{n}{\sqrt{n^2+1}}$$

for all naturals  $n$ , respectively, imply that

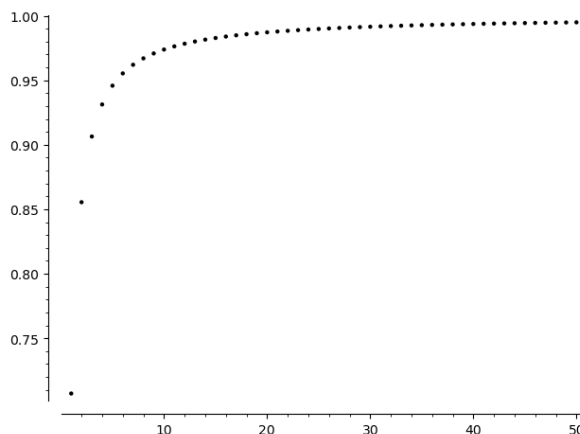
$$1 = \lim_{n \rightarrow \infty} \frac{n}{\sqrt{n^2+n}} \leq \lim_{n \rightarrow \infty} \left( \frac{1}{\sqrt{n^2+1}} + \cdots + \frac{1}{\sqrt{n^2+n}} \right) \leq \lim_{n \rightarrow \infty} \frac{n}{\sqrt{n^2+1}} = 1.$$

Thus  $\lim_{n \rightarrow \infty} \left( \frac{1}{\sqrt{n^2+1}} + \cdots + \frac{1}{\sqrt{n^2+n}} \right) = 1$ . This result can be visually confirmed in Sage by the same method used in 5.B.4.

Thus one can execute the cell

```
var("k")
p=Graphics()
for n in xrange(1, 50+1):
    p=p+points((n, sum(1/sqrt(n^2+k), k, 1, n)), color="black")
show(p)
```

This code produces the figure given below, where indeed one can observe the convergence of the given sequence to the unit.



□

**5.E.23.** Evaluate the limits  $\lim_{n \rightarrow \infty} \left( \frac{n}{n+1} \right)^n$ ,  $\lim_{n \rightarrow \infty} \left( 1 + \frac{1}{n^2} \right)^n$  and  $\lim_{n \rightarrow \infty} \left( 1 - \frac{1}{n} \right)^{n^2}$ .

○

**5.E.24.** Let  $(a_n)$  be a non-negative convergent sequence of real numbers.<sup>12</sup> Show that  $\lim_{n \rightarrow \infty} a_n = a \geq 0$ .

**Solution.** In contrast, suppose that the claim is wrong, i.e., the limit of  $(a_n)$  is a negative number,  $a < 0$ . Since  $a_n \rightarrow a$  as  $n \rightarrow \infty$  and  $\varepsilon = -a$  is a positive real number, there is a natural  $n_0$  such that  $|a_n - a| < \varepsilon$  for all  $n \geq n_0$ , that is,  $a - \varepsilon < a_n < a + \varepsilon$ , for all  $n \geq n_0$ . However,  $a = -\varepsilon$  which gives  $-2\varepsilon < a_n < 0$ , a contradiction as  $(a_n)$  consists only of non-negative terms. Thus  $a \geq 0$ . □

**5.E.25.** Let  $(a_n)$  be a sequence of real numbers and let  $(y_n)$  be a sequence of positive real numbers satisfying  $y_n \rightarrow 0$ , as  $n \rightarrow \infty$ . Suppose that for some  $N \in \mathbb{N}$ , some constant  $\mu > 0$  and some (finite) real number  $a$ , we have  $|a_n - a| < \mu y_n$  for all  $n \geq N$ . Prove that  $(a_n)$  is convergent, in particular,  $\lim_{n \rightarrow \infty} a_n = a$ .

**Solution.** The sequence  $(y_n)$  is by assumption convergent to 0, thus if  $\hat{\varepsilon} > 0$  is given there exists some  $n_0$  (depending in general on  $\hat{\varepsilon}$ ) such that  $|y_n - 0| = |y_n| = y_n < \hat{\varepsilon}$  for all  $n \geq n_0$ . By assumption  $\mu > 0$ , hence we may set  $\hat{\varepsilon} = \varepsilon/\mu$ ,

<sup>12</sup>By the term “non-negative” we mean  $a_n \geq 0$  for all naturals  $n$ .

for some  $\varepsilon > 0$ . Thus we have  $y_n < \frac{\varepsilon}{\mu}$  for all  $n \geq n_0$ . Then, for those  $n$  satisfying both  $n \geq n_0$  and  $n \geq N$  we get  $|a_n - a| \leq \mu y_n < \mu \cdot \frac{\varepsilon}{\mu} = \varepsilon$ . As  $\varepsilon > 0$  is arbitrary, the claim follows.  $\square$

**5.E.26.** (1) Let  $(a_n)$  be a sequence of positive real numbers for which the limit  $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n}$  exists and is finite real number  $\ell$ . Show that if  $\ell < 1$ , then  $\lim_{n \rightarrow \infty} a_n = 0$ .

(2) More in general, suppose that  $a_n \neq 0$  for all  $n \in \mathbb{N}$  and that the limit  $\ell = \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|$  exists. Show that if  $\ell < 1$ , then  $\lim_{n \rightarrow \infty} a_n = 0$ .

**Solution.** We will prove the first statement and leave the second one for the reader, since the method follows essentially the same ideas. We have  $a_n > 0$  for all  $n \in \mathbb{N}$ , hence the sequence  $(y_n)$  with general term  $y_n = \frac{a_{n+1}}{a_n}$  also satisfies  $y_n > 0$ , for all  $n \in \mathbb{N}$ . Also,  $y_n \rightarrow \ell$  as  $n \rightarrow \infty$  and hence the requirements of the statement in 5.E.24 are satisfied. Thus we have  $\lim_{n \rightarrow \infty} y_n = \ell \geq 0$ . Let  $b$  a real number with  $\ell < b < 1$  and set  $\varepsilon = b - \ell > 0$ . Then, there exists some  $N \in \mathbb{N}$  with  $\left| \frac{a_{n+1}}{a_n} - \ell \right| < \varepsilon$ , provided that  $n \geq N$ . This implies that  $\frac{a_{n+1}}{a_n} < \varepsilon + \ell = b - \ell + \ell = b$ , which in turn gives

$$0 < a_{n+1} < b \cdot a_n < b^2 \cdot a_{n-1} < \dots < b^{n-N+1} a_N.$$

Thus, if  $\mu := a_N b^{-N} > 0$  we obtain  $0 < a_{n+1} < \mu \cdot b^{n+1}$ , for all  $n \geq N$ . In addition, since  $0 < b < 1$  we have  $\lim_{n \rightarrow \infty} b^n = 0$ . Thus the result follows by 5.E.25.

Alternatively, the inequality  $a_{n+1} < b a_n$  for all  $n \geq N$  implies that  $a_{n+1} < a_n$ , for all  $n \geq N$  (since  $0 < b < 1$ ). Therefore, the sequence  $(a_n)$  is monotone (is decreasing starting from the  $N$ th term). Since  $(a_n)$  is also bounded, it is convergent. Let  $L = \lim_{n \rightarrow \infty} a_n$ . Then, using the relation  $a_{n+1} = y_n a_n$  we see that

$$L = \lim_{n \rightarrow \infty} a_{n+1} = \lim_{n \rightarrow \infty} (y_n a_n) = \lim_{n \rightarrow \infty} y_n \cdot \lim_{n \rightarrow \infty} a_n = \ell \cdot L \iff L(1 - \ell) = 0.$$

But  $1 - \ell > 0$ , and thus  $L = 0$ , as it is required.  $\square$

**5.E.27.** (a) Apply the statement in 5.E.25 to prove that  $\lim_{n \rightarrow \infty} \left( \frac{1}{1+na} \right) = 0$ , where  $a > 0$ .

(b) For a real number  $x$  apply the statement 5.E.26 to prove that  $\lim_{n \rightarrow \infty} \frac{x^n}{n!} = 0$ .  $\circ$

**5.E.28.** Recall that if  $(a_n)$  is a Cauchy sequence in  $\mathbb{R}$ , then the difference  $a_{n+1} - a_n$  tends to zero, i.e.,  $(a_{n+1} - a_n) \rightarrow 0$  as  $n \rightarrow \infty$ . Present a counterexample verifying that the converse of this statement is *not* in general true.

**Solution.** Consider the sequence  $(a_n = \sqrt{n})$  with  $n \in \mathbb{N}$ . Then we see that

$$a_{n+1} - a_n = \sqrt{n+1} - \sqrt{n} = \frac{(\sqrt{n+1} - \sqrt{n})(\sqrt{n+1} + \sqrt{n})}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\sqrt{n+1} + \sqrt{n}},$$

and  $\frac{1}{\sqrt{n+1} + \sqrt{n}} \rightarrow 0$ . However,  $(a_n)$  is *not* a Cauchy sequence (try to verify this claim yourself).  $\square$

**5.E.29.** Based on the definition of a Cauchy sequence, show that  $(a_n = 1/n)_{n=1}^{\infty}$  is such a sequence.

**Solution.** We need to show that for every  $\varepsilon > 0$  there exist  $N \in \mathbb{N}$  such that  $|a_n - a_m| = \left| \frac{1}{n} - \frac{1}{m} \right| < \varepsilon$  for all  $n, m \geq N$ . Indeed, for such  $n, m \geq N$  by the triangle inequality we have

$$\left| \frac{1}{n} - \frac{1}{m} \right| \leq \left| \frac{1}{n} \right| + \left| -\frac{1}{m} \right| = \frac{1}{n} + \frac{1}{m} \leq \frac{1}{N} + \frac{1}{N} = \frac{2}{N}. \quad (*)$$

Now, for  $\varepsilon > 0$  we can find non-zero  $N \in \mathbb{N}$  such that  $\frac{1}{N} < \frac{\varepsilon}{2}$ . Then by (\*) we get the result, i.e.,  $\left| \frac{1}{n} - \frac{1}{m} \right| \leq \frac{2}{N} < \varepsilon$ .  $\square$

**5.E.30.** A sequence  $(x_n)_{n=1}^{\infty}$  of real numbers is called “contractive” if there exists some real  $\alpha$  with  $0 < \alpha < 1$  such that

$$|x_{n+1} - x_n| \leq \alpha |x_n - x_{n-1}|, \quad \text{for all } n \in \mathbb{N}, n \geq 2, \quad (*)$$

Prove that any contractive sequence is Cauchy.

**Solution.** Suppose that  $(x_n)$  is a sequence of real numbers satisfying (\*) for some  $\alpha \in (0, 1)$ . By (\*) we get that

$$\begin{aligned} |x_3 - x_2| &\leq \alpha |x_2 - x_1|, \\ |x_4 - x_3| &\leq \alpha |x_3 - x_2| \leq \alpha \cdot \alpha |x_2 - x_1| = \alpha^2 |x_2 - x_1|, \\ &\dots \quad \dots \end{aligned}$$

$$|x_{n+1} - x_n| \leq \alpha |x_n - x_{n-1}| \leq \alpha^2 |x_{n-1} - x_{n-2}| \leq \dots \leq \alpha^{n-1} |x_2 - x_1|,$$

where the final relation follows by induction over  $n$ . Now, from the geometric series task 5.B.16 recall that

$$1 + \alpha + \alpha^2 + \dots + \alpha^n = \frac{1 - \alpha^{n+1}}{1 - \alpha}.$$



Combining this with our previous observation, for all  $m, n \in \mathbb{N}$  with  $m > n$  we have

$$\begin{aligned} |x_m - x_n| &\leq |x_m - x_{m-1}| + |x_{m-1} - x_{m-2}| + \cdots + |x_{n+1} - x_n| \\ &\leq \alpha^{m-2}|x_2 - x_1| + \alpha^{m-3}|x_2 - x_1| + \cdots + \alpha^{n-1}|x_2 - x_1| \\ &= |x_2 - x_1| (\alpha^{m-2} + \alpha^{m-3} + \cdots + \alpha^{n-1}) = \alpha^{n-1} (1 + \alpha + \cdots + \alpha^{m-n-1}) |x_2 - x_1| \\ &= \alpha^{n-1} \left( \frac{1 - \alpha^{m-n}}{1 - \alpha} \right) |x_2 - x_1| \leq \alpha^{n-1} \left( \frac{1}{1 - \alpha} \right) |x_2 - x_1|. \quad (\#) \end{aligned}$$

If  $x_2 = x_1$  it is easy to see that  $(x_n)$  is a Cauchy sequence, hence we may assume that  $x_2 \neq x_1$ . Recall by 5.B.3 that  $\alpha^{n-1} \rightarrow 0$  as  $n \rightarrow +\infty$ , provided that  $0 < \alpha < 1$ . Therefore, given some  $\varepsilon > 0$  we can find some positive integer  $N$  with  $|\alpha^{n-1} - 0| = |\alpha^{n-1}| < \frac{\varepsilon}{\left(\frac{1}{1-\alpha}\right) \cdot |x_2 - x_1|}$ , for all  $n \geq N$ . Then, for the same  $N$  and for  $m > n \geq N$  by (#) we see that

$$|x_m - x_n| \leq \alpha^{n-1} \left( \frac{1}{1 - \alpha} \right) |x_2 - x_1| < \frac{\varepsilon}{\left(\frac{1}{1-\alpha}\right) |x_2 - x_1|} \left( \frac{1}{1 - \alpha} \right) |x_2 - x_1| = \varepsilon.$$

Thus  $(x_n)$  is a Cauchy sequence. □

**5.E.31.** Consider the sequence  $(x_n)$  of real numbers defined recursively as follows:

$$x_1 = 1, \quad x_{n+1} = \frac{1}{4 + x_n}, \quad n \geq 1.$$

Deduce that  $(x_n)$  is a convergent sequence, by proving that it is contractive (see 5.E.30). ○

**5.E.32.** Let  $(a_n)_{n=1}^\infty$  be a bounded sequence of real numbers. Set  $m_n = \sup\{a_k : k \geq n\}$ , and  $\ell_n = \inf\{a_k : k \geq n\}$ ,  $n \in \mathbb{N}$ . Based on the monotone convergent theorem, show that both  $(m_n)_{n=1}^\infty$  and  $(\ell_n)_{n=1}^\infty$  are convergent. ○

**5.E.33. Equivalent condition for convergence.** Let  $(x_n)$  a sequence of real numbers. Is the following claim true?

$$x_n \text{ is convergent} \iff \lim_{n \rightarrow \infty} \limsup_{m \rightarrow \infty} |x_n - x_m| = 0.$$

**Solution.** The answer is *yes*. Suppose first that  $\lim_{n \rightarrow \infty} x_n = a$  for some (finite) real number  $a$ . Then,  $\lim_{m \rightarrow \infty} |x_n - x_m| = |x_n - a|$  and in turn this implies that

$$\lim_{n \rightarrow \infty} \limsup_{m \rightarrow \infty} |x_n - x_m| = \lim_{n \rightarrow \infty} |x_n - a| = 0.$$

For the opposite direction, we rely on the theory of Cauchy sequence. So, suppose that  $\lim_{n \rightarrow \infty} \limsup_{m \rightarrow \infty} |x_n - x_m| = 0$ . Then, for every  $\varepsilon > 0$  we may fix some natural  $N$  with

$$\limsup_{m \rightarrow \infty} |x_m - x_N| < \frac{\varepsilon}{2}.$$

Therefore, there exist  $m_0 \in \mathbb{N}$  with  $|x_m - x_N| < \frac{\varepsilon}{2}$  for all  $m \geq m_0$ . Combining this with the triangle inequality, for all  $m_1, m_2 \geq m_0$  we obtain

$$|x_{m_1} - x_{m_2}| \leq |x_{m_1} - x_N| + |x_{m_2} - x_N| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This show that  $(x_n)$  is a Cauchy sequence, and hence convergent, see the theorem in 5.2.3. □

**5.E.34.** Find the limes superior/inferior of the sequences  $(a_n)$ ,  $(b_n)$  and  $(c_n)$  defined below:

$$a_n = 3 + (-1)^n, \quad b_n = 2 + \frac{1}{n}, \quad c_n = \frac{4n}{n+1} \cos\left(\frac{n\pi}{2}\right), \quad (n \in \mathbb{Z}_+).$$

**Solution.** We recall that superior/inferior of a sequence  $(x_n)$  are essentially the biggest/smallest limits of all the subsequences of  $(x_n)$ . Thus, to compute them it is always sufficient to choose appropriate subsequences of  $(x_n)$ .

Let us begin with  $(a_n)$  and consider its subsequence  $(a_{2n} = 3 + (-1)^{2n})$ . This obviously tends to 4 as  $n \rightarrow +\infty$  and hence  $\limsup_{n \rightarrow \infty} a_n = 4$ . On the other hand, the subsequence  $(a_{2n+1} = 3 + (-1)^{2n+1})$  of  $(a_n)$  tends to 2 as  $n \rightarrow +\infty$  and hence  $\liminf_{n \rightarrow \infty} a_n = 2$ .

Obviously the sequence  $(b_n)$  converges to 2, so any subsequence of  $(b_n)$  will have the same limit and it follows that  $\limsup_{n \rightarrow \infty} b_n = 2 = \liminf_{n \rightarrow \infty} b_n$ .

For  $(c_n)$  we have the expression  $c_n = x_n \cdot y_n$ , where  $x_n = \frac{4n}{n+1}$  and  $y_n = \cos\left(\frac{n\pi}{2}\right)$ . Obviously,  $\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} \frac{4}{1+\frac{1}{n}} = 4$  and so we may focus on  $(y_n)$ . We consider the following four subsequences:

- The subsequence of  $(y_n)$  defined by  $y_{4n} = \cos\left(\frac{4n\pi}{2}\right) = \cos(2n\pi) = 1$ , which satisfies  $\lim_{n \rightarrow \infty} y_{4n} = 1$ . Thus, the

subsequence  $(c_{4n})$  of  $(c_n)$  defined by  $c_{4n} = x_{4n} \cdot y_{4n}$  will tend to  $4 \cdot 1 = 4$ .

• The subsequence of  $(y_n)$  defined by  $y_{4n+1} = \cos(\frac{(4n+1)\pi}{2}) = \cos(2n\pi + \frac{\pi}{2}) = \cos(\pi/2) = 0$ , satisfies  $\lim_{n \rightarrow \infty} y_{4n+1} = 0$ . Hence in this case the subsequence  $(c_{4n+1})$  of  $(c_n)$  defined by  $c_{4n+1} = x_{4n+1} \cdot y_{4n+1}$  tends to  $4 \cdot 0 = 0$ .

• The subsequence of  $(y_n)$  defined by  $y_{4n+2} = \cos(\frac{(4n+2)\pi}{2}) = \cos(2n\pi + \pi) = \cos(\pi) = -1$ , satisfies  $\lim_{n \rightarrow \infty} y_{4n+2} = -1$ . Thus, the subsequence  $(c_{4n+2})$  of  $(c_n)$  defined by  $c_{4n+2} = x_{4n+2} \cdot y_{4n+2}$  tends to  $4 \cdot -1 = -4$ .

• The subsequence of  $(y_n)$  defined by  $y_{4n+3} = \cos(\frac{(4n+3)\pi}{2}) = \cos(2n\pi + \frac{3\pi}{2}) = \cos(\frac{3\pi}{2}) = 0$  tends to 0 and hence the subsequence  $(c_{4n+3})$  of  $(c_n)$  tends to 0 as well.

To summarize, the largest limit of the subsequences  $(c_{4n})$ ,  $(c_{4n+1})$ ,  $(c_{4n+2})$  and  $(c_{4n+3})$  is 4 and the smallest one is  $-4$ . This means that  $\limsup_{n \rightarrow \infty} c_n = 4$  and  $\liminf_{n \rightarrow \infty} c_n = -4$ , respectively.  $\square$

**5.E.35.** Based on the result from 5.B.18 on the Euler number, compute  $\limsup_{n \rightarrow \infty} a_n$  and  $\liminf_{n \rightarrow \infty} a_n$ , where  $(a_n)_{n=1}^{\infty}$  is the sequence defined by  $a_n = \left(1 + \frac{(-1)^n}{n}\right)^n$ , for  $n \in \mathbb{Z}_+$ .

**Solution.** Consider a sequence  $(a_n)$  which its definition involves the expression  $(-1)^n$  (or  $(-1)^{n+1}$  or  $(-1)^{n-1}$ ). Recall that in this case in order to compute the superior/inferior of  $(a_n)$  we can consider the subsequences determined for  $2n$  and  $2n + 1$ , respectively. Hence for the specific task let us consider first the subsequence of  $(a_n)$  defined by  $a_{2n} = \left(1 + \frac{1}{2n}\right)^{2n}$ . We have  $2n \rightarrow +\infty$  as  $n \rightarrow +\infty$ , and hence by 5.B.18 we see that  $\lim_{n \rightarrow +\infty} a_{2n} = e$ . On the other hand, the subsequence of  $(a_n)$  defined by  $a_{2n+1} = \left(1 + \frac{(-1)^{2n+1}}{2n+1}\right)^{2n+1}$  satisfies  $a_{2n+1} = \left(1 - \frac{1}{2n+1}\right)^{2n+1}$  and a combination of 5.B.18 with the exercise 5.E.23 above, gives  $\lim_{n \rightarrow +\infty} a_{2n+1} = e^{-1}$ . As a conclusion we get that  $\limsup_{n \rightarrow \infty} a_n = e$  and  $\liminf_{n \rightarrow \infty} a_n = e^{-1}$ .  $\square$

*Let us now present a few extra tasks on the topological notions that we met so far in Chapter 5.*

**5.E.36.** Is some of the sets  $\mathbb{N}$  or  $\mathbb{Q}$  an open or a closed subset of  $\mathbb{R}$ ?  $\circ$

**5.E.37.** (a) Given the sets  $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ ,  $\mathbb{R} \setminus \mathbb{Q}$ ,  $\mathbb{R} \setminus \mathbb{Z}$ , and  $(1, 2) \cup \{5\}$ , locate the open one.  $\circ$

(b) Similarly, given the sets  $\mathbb{R} \setminus \mathbb{Q}$ ,  $\mathbb{R} \setminus \mathbb{Z}$  and  $[1, 2] \cup \{5\}$ , locate the closed one.  $\circ$

**5.E.38.** (a) Consider the set  $A = \bigcap_{n \in \mathbb{N}^*} \left(-\frac{1}{n}, 1 + \frac{1}{n}\right)$ , where as usual  $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ . Show that  $A = [0, 1]$ .

(b) Is the set  $B = \bigcup_{n \in \mathbb{N}^*} \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$  an open or a closed subset of  $\mathbb{R}$ ?

**Solution.** (a) Clearly,  $[0, 1] \subset A$  and we need to prove the opposite inclusion. So assume that  $x \in A$ , that is,  $-\frac{1}{n} < x < 1 + \frac{1}{n}$  for all  $n \in \mathbb{Z}_+$ . Then we have  $x \geq \sup\{-\frac{1}{n} : n \in \mathbb{Z}_+\}$  and  $x \leq \inf\{1 + \frac{1}{n} : n \in \mathbb{Z}_+\}$ . However,  $\sup\{-\frac{1}{n} : n \in \mathbb{Z}_+\} = 0$  and  $\inf\{1 + \frac{1}{n} : n \in \mathbb{Z}_+\} = 1$ , which means that  $x \in [0, 1]$ .

(b) This can be treated similarly, in particular one can show that

$$B = \bigcup_{n \in \mathbb{N}^*} \left[\frac{1}{n}, 1 - \frac{1}{n}\right] = (0, 1),$$

which we leave for practice. Hence  $B$  is open and this shows that the union of infinitely many closed sets is *not* necessary closed.  $\square$

**5.E.39. Compact sets.** Suppose that we define the compactness of a subset  $A \subset \mathbb{R}$  via the Bolzano-Weierstrass theorem (see (4) in Theorem 5.2.8), that is,  $A$  is said to be compact if for any sequence  $(x_n)$  in  $A$  there is a convergent subsequence  $(x_{n_k})$  of  $(x_n)$  whose limit belongs to  $A$ . Under this definition, show that a compact set  $A$  is closed and bounded.

**Solution.** If  $A$  is the empty set, then the result clearly holds, and hence assume that  $A \neq \emptyset$  and that  $A$  is compact. According to our new definition, this means that for any sequence  $(x_n)$  in  $A$  there is a convergent subsequence  $(x_{n_k})$  of  $(x_n)$  whose limit, say  $\ell$ , belongs to  $A$ . Hence if  $(x_n)$  is convergent, then we should have  $\lim_{n \rightarrow \infty} x_n = \lim_{k \rightarrow \infty} x_{n_k} = \ell \in A$ . This shows that  $A$  is closed and it remains to show that  $A$  is bounded. For each  $n \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$  consider the open sets  $C_n = (-n, n)$ . Obviously,

$$\bigcup_{n \in \mathbb{N}^*} C_n = \mathbb{R}, \quad \text{hence} \quad A \subseteq \bigcup_{n \in \mathbb{N}^*} C_n.$$

Since each  $C_n$  is also open, this means that the family  $\{C_n : n \in \mathbb{N}^*\}$  is an open cover of  $A$ . However,  $A$  is compact and by (5) in Theorem 5.2.8 each of its open covers contains a finite subcover of  $A$ . Thus there exists  $m \in \mathbb{N}^*$  such that

$$A \subseteq \bigcup_{n=1}^m C_n$$

with  $\bigcup_{n=1}^m C_n = (-m, m)$ . This shows that  $A$  is contained in the bounded interval  $(-m, m)$ , therefore  $A$  is bounded. For instance, any interval of the form  $[a, b] \subset \mathbb{R}$  with  $a < b$ ,  $a, b \in \mathbb{R}$ , is closed (and bounded) and hence compact.  $\square$

**5.E.40. An open cover.** Find an open cover, a subcover and a finite subcover of  $A := \{x \in \mathbb{R} : 0 \leq x \leq 1\} = [0, 1]$ .

**Solution.** An open cover of  $A$  is given by the family  $\mathcal{A} := \{\mathcal{A}_a = (a - \varepsilon, a + \varepsilon)\}$  where  $a \in A$  and  $\varepsilon > 0$ . As a subcover, for the same  $\varepsilon$  take the subfamily  $\{\mathcal{A}_b = (b - \varepsilon, b + \varepsilon)\}$ , where  $b \in \{x \in \mathbb{Q} : 0 \leq x \leq 1\}$ . Notice however that this is *not* a finite subcover. To obtain a finite subcover, consider the family

$$\{\mathcal{A}_{0 \cdot \varepsilon}, \mathcal{A}_{1 \cdot \varepsilon}, \mathcal{A}_{2 \cdot \varepsilon}, \dots, \mathcal{A}_{n_0 \cdot \varepsilon}\} = \{\mathcal{A}_0, \mathcal{A}_\varepsilon, \mathcal{A}_{2\varepsilon}, \dots, \mathcal{A}_{n_0\varepsilon}\},$$

where  $n_0 \in \mathbb{N}$  is the largest natural with  $n_0\varepsilon < 1$ . Then this is finite subfamily of  $\{\mathcal{A}_a = (a - \varepsilon, a + \varepsilon) : a \in A\}$  (every set in this family is obviously a member of  $\mathcal{A}$ ), and moreover it covers  $A = [0, 1]$ , i.e.,  $A \subseteq \mathcal{A}_0 \cup \mathcal{A}_\varepsilon \cup \mathcal{A}_{2\varepsilon} \cup \dots \cup \mathcal{A}_{n_0\varepsilon}$ .  $\square$

**5.E.41.** Show that the collection of open intervals  $\mathcal{A} = \{\mathcal{A}_k = (k - 1, k + 1)\}_{k \in \mathbb{N}^*}$  is an open cover of  $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ . Deduce however that  $\mathbb{N}^*$  is not a compact subset of  $\mathbb{R}$ .

**5.E.42.** Indeed, we see that

$$\bigcup_{k=1}^{\infty} \mathcal{A}_k = (0, +\infty)$$

and  $\mathbb{N}^* \subset (0, +\infty)$ . Observe however that the family  $\{\mathcal{A}_k = (k - 1, k + 1)\}_{k \in \mathbb{N}^*}$  is not a cover of  $\mathbb{N}$ , since its union does not contain 0.<sup>13</sup> On the other hand we see that there is no finite subfamily  $\{\mathcal{A}_{i_1}, \dots, \mathcal{A}_{i_n}\}$  of  $\mathcal{A}$  which covers  $\mathbb{N}^*$ . For, if  $\alpha = \max\{i_1, \dots, i_n\}$  then we see that

$$\bigcup_{k=1}^n \mathcal{A}_{i_k} \subset (0, \alpha + 1),$$

and it is obvious that this union does not contain all naturals. Hence  $\mathbb{N}^*$  is a non-compact set.  $\circ$

**5.E.43.** Suppose that  $A \subset \mathbb{R}$  is a non-empty compact subset of  $\mathbb{R}$  and let  $\varepsilon > 0$ . Specify a finite subset  $B \subset \mathbb{R}$  such that  $\min\{|x - y| : y \in B\} < \varepsilon$  for all  $x \in A$ .

**Solution.** It is easy to see that for any  $\varepsilon > 0$  the set  $\{(x - \varepsilon, x + \varepsilon) : x \in A\}$  is an open cover of  $A$ . On the other hand,  $A$  is by assumption compact. Hence there exist some positive integer  $n$  and  $x_1, \dots, x_n$ , all points of  $A$ , such that

$$A \subset \bigcup_{i=1}^n (x_i - \varepsilon, x_i + \varepsilon).$$

Hence a finite subset  $B \subset \mathbb{R}$  with the required properties is given by these points, i.e.,  $B = \{x_1, \dots, x_n\}$ .  $\square$

*The theory of limits often combines with the theory of matrices, and can be used to solve tasks related to linear algebra.*



For instance, let  $(A_n)_{n=1}^{\infty}$  be a sequence of  $2 \times 2$  matrices, with  $A_n = \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix}$ , where  $(a_n)_{n=1}^{\infty}$ ,  $(b_n)_{n=1}^{\infty}$ ,  $(c_n)_{n=1}^{\infty}$  and  $(d_n)_{n=1}^{\infty}$  are sequences of real (or complex) numbers. When  $a_n \rightarrow a$ ,  $b_n \rightarrow b$ ,  $c_n \rightarrow c$  and  $d_n \rightarrow d$  for some real numbers  $a, b, c, d$  we say that  $(A_n)_{n=1}^{\infty}$  converges to the matrix  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ , i.e.,  $\lim_{n \rightarrow \infty} A_n = A$ .

In this case  $A$  referred to as the “limit matrix” of  $(A_n)_{n=1}^{\infty}$ . Let us illustrate such an example.

**5.E.44.** Given a positive real number  $\varphi$ , consider the matrix sequence  $(A_n)$  whose general term is given by

$$A_n = \begin{pmatrix} 1 & -\frac{\varphi}{n} \\ \frac{\varphi}{n} & 1 \end{pmatrix}^n, \quad n \in \mathbb{Z}_+.$$

Based on the relation  $\lim_{n \rightarrow \infty} \left(1 + \frac{z}{n}\right)^n = e^z$  for all  $z \in \mathbb{C}$ , prove that the limit matrix  $A$  of  $(A_n)$  exists and describes a rotation in the plane by angle  $\varphi$ , i.e.,

$$A = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{pmatrix}.$$

<sup>13</sup>Recall that till Chapter 11 we assume that  $\mathbb{N}$  contains 0.

**Solution.** Let us set  $\frac{\varphi}{n} = \tan(\theta_n)$ . Based on basic matrix calculus we can then express the matrix sequence  $(A_n)$  as follows:

$$\begin{aligned} A_n &= \begin{pmatrix} 1 & -\frac{\varphi}{n} \\ \frac{\varphi}{n} & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & -\tan(\theta_n) \\ \tan(\theta_n) & 1 \end{pmatrix}^n = \frac{1}{\cos^n(\theta_n)} \begin{pmatrix} \cos(\theta_n) & -\sin(\theta_n) \\ \sin(\theta_n) & \cos(\theta_n) \end{pmatrix}^n \\ &= \frac{1}{\cos^n(\theta_n)} \begin{pmatrix} \cos(n\theta_n) & -\sin(n\theta_n) \\ \sin(n\theta_n) & \cos(n\theta_n) \end{pmatrix} = \begin{pmatrix} \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} & -\frac{\sin(n\theta_n)}{\cos^n(\theta_n)} \\ \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} & \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} \end{pmatrix}. \end{aligned}$$

Notice above we used the relations

$$\begin{pmatrix} 1 & -\tan(\theta_n) \\ \tan(\theta_n) & 1 \end{pmatrix} = \frac{1}{\cos(\theta_n)} \begin{pmatrix} \cos(\theta_n) & -\sin(\theta_n) \\ \sin(\theta_n) & \cos(\theta_n) \end{pmatrix}, \quad \begin{pmatrix} \cos(\theta_n) & -\sin(\theta_n) \\ \sin(\theta_n) & \cos(\theta_n) \end{pmatrix}^n = \begin{pmatrix} \cos(n\theta_n) & -\sin(n\theta_n) \\ \sin(n\theta_n) & \cos(n\theta_n) \end{pmatrix}.$$

Although the first is obvious, you may like to confirm the second one by induction. Hence we can write

$$\lim_{n \rightarrow \infty} A_n = \begin{pmatrix} \lim_{n \rightarrow \infty} \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} & -\lim_{n \rightarrow \infty} \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} \\ \lim_{n \rightarrow \infty} \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} & \lim_{n \rightarrow \infty} \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} \end{pmatrix}, \quad (*)$$

and  $A = \lim_{n \rightarrow \infty} A_n$  will exist if and only if the limits inside the matrix in  $(*)$  exist. To examine these limits we rely on de Moivre's theorem, as follows:

$$\frac{\cos(n\theta_n)}{\cos^n(\theta_n)} + i \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} = \frac{\cos(n\theta_n) + i \sin(n\theta_n)}{\cos^n(\theta_n)} = \left( \frac{\cos(\theta_n) + i \sin(\theta_n)}{\cos(\theta_n)} \right)^n = \left( 1 + i \tan(\theta_n) \right)^n = \left( 1 + \frac{i\varphi}{n} \right)^n.$$

Recalling that  $\lim_{n \rightarrow \infty} \left( 1 + \frac{i\varphi}{n} \right)^n = e^{i\varphi} = \cos(\varphi) + i \sin(\varphi)$ , we deduce that

$$\lim_{n \rightarrow \infty} \left( \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} + i \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} \right) = \cos(\varphi) + i \sin(\varphi) \iff \lim_{n \rightarrow \infty} \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} + i \lim_{n \rightarrow \infty} \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} = \cos(\varphi) + i \sin(\varphi).$$

Comparing the real and imaginary parts, we obtain  $\lim_{n \rightarrow \infty} \frac{\cos(n\theta_n)}{\cos^n(\theta_n)} = \cos(\varphi)$ ,  $\lim_{n \rightarrow \infty} \frac{\sin(n\theta_n)}{\cos^n(\theta_n)} = \sin(\varphi)$ , and the result follows by  $(*)$ .  $\square$

Let's now discuss additional content concerning the limits of functions and the concept of continuity. This will involve various computational and theoretical examples, as well as notable applications of the intermediate value theorem. However, we will first address tasks involving sequences defined recursively, such as in the problem 5.E.31 above.



Let  $f : A \subset \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function defined on a subset  $A$  of real numbers. Suppose that exists a sequence  $(x_n)$  in  $A$  satisfying  $x_{n+1} = f(x_n)$ , for all  $n$ . Moreover, assume that the  $\lim_{n \rightarrow \infty} x_n$  exists, and that is equal to some (limit point)  $a \in A$ . In this situation it is not hard to see that  $\lim_{n \rightarrow \infty} f(x_n) = f(a)$  and  $f(a) = a$ . This shows how useful can be continuous functions and provides a very common way to define sequences, that is, by iterating some function  $f$  (recursive definition).

**5.E.45. Iterations.** Consider the sequence  $(a_n)_{n=1}^{\infty}$  with  $a_1 = 0$ ,  $a_2 = 1$ ,  $a_3 = \sqrt{3}$ ,  $\dots$ ,  $a_n = \sqrt{1 + 2a_{n-1}}$ .

(a) Show that  $(a_n)_{n=1}^{\infty}$  is strictly increasing.

(b) Is this a convergent sequence? In the positive case, find the limit  $\lim_{n \rightarrow \infty} a_n$ .

(c) Use Sage to plot some terms of  $(a_n)$  and obtain a graphical confirmation of your conclusion in (b).

**Solution.** We should prove that  $a_{n+1} > a_n$  for all  $n$ . Observe that  $a_2 = 1 > 0 = a_1$ . Assuming that  $a_n > a_{n-1}$  for  $n \geq 2$  and proceeding by induction over  $n$  one gets  $a_{n+1} = \sqrt{1 + 2a_n} > \sqrt{1 + 2a_{n-1}} = a_n$ . Using mathematical induction over  $n$  we can also show that  $(a_n)$  is bounded above, that is,  $a_n < 3$  for all naturals  $n \geq 1$ . This is because  $a_1 = 0 < 3$  and hence assuming  $a_n < 3$  for  $n \geq 2$ , we can arrive to the claim:  $a_{n+1} = \sqrt{1 + 2a_n} < \sqrt{1 + 2 \cdot 3} < \sqrt{9} = 3$ . Since the sequence  $(a_n)$  is strictly increasing and bounded above, by the monotone sequence theorem it should be convergent.

To determine its limit we rely on the continuity of the square root function  $\sqrt{x}$ , as follows: Suppose that  $a = \lim_{n \rightarrow \infty} a_n$ . Then  $a$  must satisfy the relation  $a = \sqrt{1 + 2a}$ , thus  $a^2 = 1 + 2a$  with solutions  $a = 1 \pm \sqrt{2}$ . However,  $a_n \geq 0$  for all  $n$ , from where one can neglect the negative solution, that is,  $\lim_{n \rightarrow \infty} a_n = 1 + \sqrt{2}$ .

Let us now use Sage to plot some terms of  $(a_n)$ . To do so we will first use the `def` command to introduce our sequence. This can be done as follows:

```

def a(n, D={}):
    if n in D.keys():
        return D[n]
    if (n==1):
        sol = 0
    else:
        sol = sqrt(1+2*a(n-1))
    D[n] = sol
    return sol
    
```

In this way to test different values of  $(a_n)$  one can simply type

```
a(2).n(); a(3).n(); a(10).n()
```

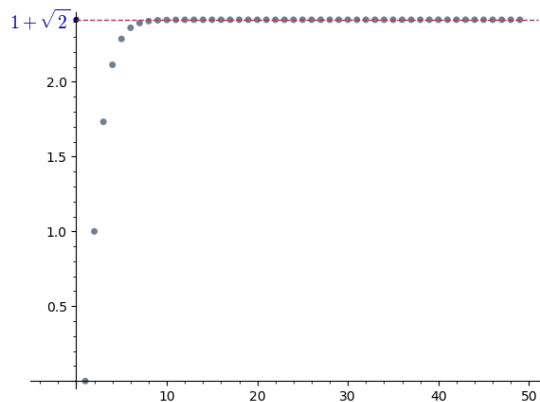
etc. Notice however that the definition of  $(a_n)$  via the `def` method, does not allow us to compute the limit of  $(a_n)$ , which means that typing `lim(a(n), n = oo)` returns an error (even if we add before the `def` command the variable  $n$  as symbolic).

Now we can use this routine to plot  $(a_n)$ . Together we will sketch the line  $y = 1 + \sqrt{2}$  so that the visualization of our conclusion about the limit to be easier.

```

N=50
pts = [(n, a(n)) for n in range(1, N)]
p=points(pts, color="slategray", size=25)
p+=line([(0, 1+sqrt(2)), (51, 1+sqrt(2))], rgbcolor=(0.7,0.2,0.3), linestyle="--")
p+=point([0, 1+sqrt(2)], size=20, color="darkblue")
p+=text(r"$1+\sqrt{2}$", (-4, 1+sqrt(2)), color="darkblue", fontsize="14")
p.show(figsize=6)
    
```

The figure that Sage returns is here:



□

**5.E.46.** Consider the sequence  $(a_n)_{n=1}^{\infty}$  defined by the relation  $a_{n+1} = \frac{1}{2} \left( a_n + \frac{a^2}{a_n} \right)$  for all positive integers  $n$ , with  $a_1 > a$ , for some  $a > 0$ . Show that  $\lim_{n \rightarrow +\infty} a_n = a$ .

**Solution.** For all naturals  $n > 1$  by the AM-GM-inequality presented in Chapter 1, we have

$$a_n = \frac{1}{2} \left( a_{n-1} + \frac{a^2}{a_{n-1}} \right) \geq \sqrt{a_{n-1} \cdot \frac{a^2}{a_{n-1}}} = a > 0.$$

Thus, together with the condition  $a_1 > a$  this gives  $a_n \geq a$  for all  $n$  and hence  $(a_n)$  is bounded. It is also decreasing, since

$$a_{n+1} - a_n = \frac{a^2 - a_n^2}{2a_n} \leq 0. \quad (b)$$

One may like to confirm the equality appearing in (b) also via Sage, which can be done quickly by the block

```

var("n, c"); function("a")(n)
bool((1/2)*(a(n)+(c^2/a(n)))-a(n)==(c^2-a(n)*a(n))/(2*a(n)))
    
```

In this way one proves that  $(a_n)$  is bounded and monotone, hence its limit  $\lim_{n \rightarrow +\infty} a_n$  exists and equals to  $\inf\{a_n : n \in \mathbb{N}\}$ .

Let us set  $b = \inf\{a_n : n \in \mathbb{N}\}$ . Then we should have

$$b = \frac{1}{2} \left( b + \frac{a^2}{b} \right),$$

which can be equivalently expressed as  $b^2 = a^2$ . However, by assumption  $a > 0$  and thus we get  $b = a$ . □

**5.E.47.** Consider the sequence  $(a_n)_{n=1}^{\infty}$  with  $a_1 = \sqrt{2}$  and  $a_n = \sqrt{2 + a_{n-1}}$  for all naturals  $n$  with  $n \geq 2$ . Show that  $(a_n)$  is convergent, and in particular compute  $\lim_{n \rightarrow +\infty} a_n$ . ○

**5.E.48.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function satisfying the relation  $x^2 f(x) = 1 - \cos(2x)$ , for all  $x \in \mathbb{R}$ . Using the result of 5.B.31, find the explicit form of  $f$ .

**Solution.** Obviously, for  $x \neq 0$  the function at hand satisfies  $f(x) = \frac{1 - \cos(2x)}{x^2}$ . For  $x = 0$ , by the continuity of  $f$  we should have  $f(0) = \lim_{x \rightarrow 0} f(x)$ . To compute this limit one can use the known identity  $\sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2}$  (half-angle formula), which can be rephrased as  $2 \sin^2(\theta) = 1 - \cos(2\theta)$ . This is a very useful trigonometric identity, which for our case applies as follows:

$$f(0) = \lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} \frac{1 - \cos(2x)}{x^2} = 2 \lim_{x \rightarrow 0} \frac{\sin^2(x)}{x^2} = 2 \left( \lim_{x \rightarrow 0} \frac{\sin(x)}{x} \right)^2 = 2 \cdot 1^2 = 2.$$

Therefore one can now present the full type of  $f$ , namely

$$f(x) = \begin{cases} \frac{1 - \cos(2x)}{x^2}, & \text{for } x \neq 0, \\ 2, & \text{for } x = 0. \end{cases}$$

□

**5.E.49.** Using the result of 5.B.31, when adequate, evaluate the following limits:

$$\lim_{x \rightarrow 0} \frac{\sin^2(x)}{x}, \quad \lim_{x \rightarrow 0} \frac{x}{\sin^2(x)}, \quad \lim_{x \rightarrow 0} \frac{\arcsin(x)}{x}, \quad \lim_{x \rightarrow 0} \frac{3 \tan^2(x)}{5x^2}, \quad \lim_{x \rightarrow 0} \frac{\sin(3x)}{\sin(5x)}, \quad \lim_{x \rightarrow 0} \frac{\tan(3x)}{\sin(5x)}.$$

○

**5.E.50.** Confirm that

$$\lim_{x \rightarrow 0} \frac{e^{5x} - e^{2x}}{x} = 3 = \lim_{x \rightarrow 0} \frac{e^{5x} - e^{-x}}{\sin(2x)}.$$

(Hint: Recall that  $\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1$ ). Next use Sage to sketch the graphs of the involved functions. ○

**5.E.51.** Compute the limit  $\lim_{x \rightarrow +\infty} \left( \frac{3x+1}{3x-2} \right)^{4x}$  with the aid of the relation  $\lim_{x \rightarrow +\infty} \left( 1 + \frac{1}{x} \right)^x = e$ . Next verify your answer via Sage.

**Solution.** Based on the properties of powers, we see that

$$\left( \frac{3x+1}{3x-2} \right)^{4x} = \frac{[3x(1 + \frac{1}{3x})]^{4x}}{[3x(1 - \frac{2}{3x})]^{4x}} = \frac{(1 + \frac{1}{3x})^{4x}}{(1 - \frac{2}{3x})^{4x}} = \frac{\left[ \left( 1 + \frac{1}{3x} \right)^{3x} \right]^{\frac{4}{3}}}{\left[ \left( 1 + \frac{-2}{3x} \right)^{3x} \right]^{\frac{4}{3}}} = \frac{\left[ \left( 1 + \frac{1}{3x} \right)^{3x} \right]^{\frac{4}{3}}}{\left[ \left( 1 + \frac{-\frac{2}{3}x}{x} \right)^{-\frac{3}{2}x} \right]^{\frac{4}{3}}}.$$

Thus

$$\lim_{x \rightarrow +\infty} \left( \frac{3x+1}{3x-2} \right)^{4x} = \frac{\lim_{x \rightarrow +\infty} \left[ \left( 1 + \frac{1}{3x} \right)^{3x} \right]^{\frac{4}{3}}}{\lim_{x \rightarrow +\infty} \left[ \left( 1 + \frac{-\frac{2}{3}x}{x} \right)^{-\frac{3}{2}x} \right]^{\frac{4}{3}}} = \frac{\left[ \lim_{x \rightarrow +\infty} \left( 1 + \frac{1}{3x} \right)^{3x} \right]^{\frac{4}{3}}}{\left[ \lim_{x \rightarrow +\infty} \left( 1 + \frac{-\frac{2}{3}x}{x} \right)^{-\frac{3}{2}x} \right]^{\frac{4}{3}}} = \frac{e^{\frac{4}{3}}}{e^{-\frac{8}{3}}} = e^{\frac{4}{3} + \frac{8}{3}} = e^4.$$

For a confirmation via Sage just give the cell `limit(((3 * x + 1)/(3 * x - 2))**(4 * x), x = oo)`. □

**5.E.52.** Evaluate the limits  $\lim_{x \rightarrow +\infty} \left( 2 + \frac{1}{x} \right)^{\frac{1}{x}}$ ,  $\lim_{x \rightarrow +\infty} x^{-x}$  and  $\lim_{x \rightarrow 0} e^{\frac{1}{x}}$ . ○

**5.E.53.** Evaluate the following limits:

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x^3}, \quad \lim_{x \rightarrow +\infty} \left( \sqrt{x^2 + x} - x \right), \quad \lim_{x \rightarrow +\infty} \left( x \sqrt{1 + x^2} - x^2 \right), \quad \lim_{x \rightarrow 0^-} \frac{\sqrt{1 + \tan(x)} - \sqrt{1 - \tan(x)}}{\sin(x)}.$$

**5.E.54.** Use the binomial theorem to compute the limit  $\lim_{x \rightarrow 0} \frac{(1 + 2nx)^n - (1 + nx)^{2n}}{x^2}$ , for all  $n \in \mathbb{N}^*$ . Next confirm your computation via Sage. ○

**5.E.55.** Exam the convergence of the sequence  $(a_n)$  with general term ○

$$a_n = \sum_{k=1}^n \frac{\cos(k)}{2^k}, \quad n = 1, 2, \dots$$

**5.E.56.** Compute the limit  $\lim_{n \rightarrow +\infty} \frac{\lfloor xn \rfloor}{n}$ , where  $x \in \mathbb{R}$  and  $\lfloor \cdot \rfloor$  is the floor function introduced in 5.B.48. ○

**Solution.** Recall from 5.B.48 that for any  $x \in \mathbb{R}$  we have  $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$ . Since in general we have  $nx \in \mathbb{R}$  for all  $n \in \mathbb{Z}$  and  $x \in \mathbb{R}$  we also get  $\lfloor nx \rfloor \leq nx < \lfloor nx \rfloor + 1$ , which we may rewrite as  $0 \leq x - \frac{\lfloor xn \rfloor}{n} < \frac{1}{n}$ , for all  $n \neq 0$ . Combining this with the squeeze theorem we deduce that  $\lim_{n \rightarrow +\infty} \frac{\lfloor xn \rfloor}{n} = x$ , for all  $x \in \mathbb{R}$ . □

**5.E.57. The ceiling function and the fractional part function.** (1) Use the command `ceil` in Sage to compute  $\lceil -1.4 \rceil$ ,  $\lceil -0.5 \rceil$ ,  $\lceil 2 \rceil$ ,  $\lceil 2.1 \rceil$ ,  $\lceil \sqrt{7} \rceil$ ,  $\lceil \pi \rceil$  and  $\lceil 9/2 \rceil$ .  
 (2) Construct in Sage the fractional part function and then compute its evaluation on the values mentioned in (1).  
 (3) Show that  $\lceil x \rceil = -\lfloor -x \rfloor$ , for all  $x \in \mathbb{R}$ .  
 (4) If  $x = n + d$ , with  $n \in \mathbb{Z}$  and  $0 \leq d < 1$ , show that  $n = \lfloor x \rfloor$  and  $d = \{x\}$ .  
 (5) Show that the fractional part function is discontinuous at any integer.  
 (6) Use Sage to sketch the graph of the ceiling function for  $-5 \leq x \leq 5$  and of the fractional part function for  $0 \leq x \leq 5$ .

**Solution.** (1) It is not hard to compute

$$\lceil -1.4 \rceil = -1, \quad \lceil -0.5 \rceil = 0, \quad \lceil 2 \rceil = 2, \quad \lceil 2.1 \rceil = 3 = \lceil \sqrt{7} \rceil, \quad \lceil \pi \rceil = 4, \quad \lceil 9/2 \rceil = 5.$$

A confirmation in Sage occurs by the command `ceil`, as follows:

```
print(ceil(-1.4)); print(ceil(-0.5)); print(ceil(2)); print(ceil(2.1))
print(ceil(sqrt(7))); print(ceil(pi)); print(ceil(9/2))
```

(2) In Sage to introduce the fractional part function  $x \mapsto \{x\} = x - \lfloor x \rfloor$ , use the floor function `floor` and type the syntax: `fractal(x) = x - floor(x)` Then, to compute the required evaluations type `print(fractal(-1.4))`, `print(fractal(-0.5))`, etc. Notice that Sage has a build-in function for the fractional part function, called `frac` but this agrees with ours only for positive reals. In this way you can quickly verify the following answers:

$$\begin{aligned} \{-1.4\} &= -1.4 - \lfloor -1.4 \rfloor = -1.4 - (-2) = 0.6, & \{-0.5\} &= -0.5 - \lfloor -0.5 \rfloor = -0.5 - (-1) = 0.5, \\ \{2\} &= 2 - \lfloor 2 \rfloor = 2 - 2 = 0, & \{2.1\} &= 2.1 - \lfloor 2.1 \rfloor = 2.1 - 2 = 0.1, \\ \{\sqrt{7}\} &= \sqrt{7} - \lfloor \sqrt{7} \rfloor \approx 2.645752 - 2 \approx 0.645752, & \{\pi\} &= \pi - \lfloor \pi \rfloor \approx 0.141593, & \{9/2\} &= 4.5 - \lfloor 4.5 \rfloor = 0.5. \end{aligned}$$

(3) We leave this to the reader.

(4) Consider a real  $x$  such that  $x = n + d$  with  $n \in \mathbb{Z}$  and  $0 \leq d < 1$ . By (2) in 5.B.48 we know that  $\lfloor x + n \rfloor = \lfloor x \rfloor + n$  for all  $x \in \mathbb{R}$  and  $n \in \mathbb{Z}$ . Thus we have  $\lfloor x \rfloor = \lfloor n + d \rfloor = n + \lfloor d \rfloor = n$ , since  $\lfloor d \rfloor = 0$ . Consequently,

$$x = n + d = \lfloor x \rfloor + d = \lfloor x \rfloor + \{x\},$$

that is,  $\{x\} = d$ . This proves (4).

(5) Let  $n \in \mathbb{Z}$  be an integer. We see that

$$\lim_{x \rightarrow n^+} \{x\} = \lim_{h \rightarrow 0} \{n + h\} = \lim_{h \rightarrow 0} (n + h - \lfloor n + h \rfloor) = \lim_{h \rightarrow 0} (n + h) - \lim_{h \rightarrow 0} (\lfloor n + h \rfloor) = n - \lim_{h \rightarrow 0} (n + \lfloor h \rfloor) = n - n = 0.$$

On the other hand, we compute

$$\lim_{x \rightarrow n^-} \{x\} = \lim_{h \rightarrow 0} \{n - h\} = \lim_{h \rightarrow 0} (n - h - \lfloor n - h \rfloor) = \lim_{h \rightarrow 0} (n - h) - \lim_{h \rightarrow 0} (\lfloor n - h \rfloor) = n - (n - 1) = 1.$$

It follows that the function  $x \mapsto \{x\}$  is not continuous at any integer (see also its graph below).

(6) For this task and for the ceiling function you may proceed in analogous way with the floor function, presented in 5.B.48. This method includes the jump discontinuities, and is encoded by the following block:

```
g=ceil(x)
p=plot(g, x, -5, 5, ticks=[1,1])
for x in [-5..5] :
```

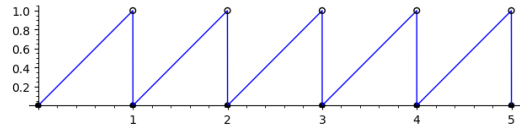
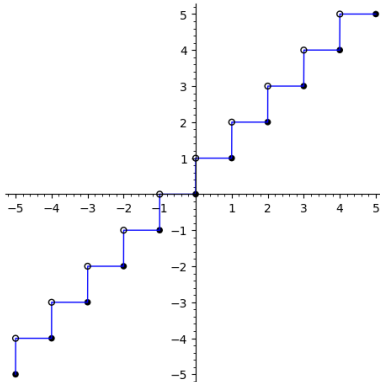


```
p+=point([x,x], size=30, color="black")
for x in [-4..5]:
    p+=circle((x-1,x),0.08, color="black")
show(p)
```

In a similar way we can sketch the graph of the fractional part function (together with the jump discontinuities).

```
fractal(x)=x-floor(x)
q=plot(fractal, x, 0, 5)
for x in [0..5]:
    q+=point([x,0], size=30, color="black")
for x in [1..5]:
    q+=circle((x,1),0.03, color="black")
show(q)
```

Let us present the figures that Sage constructs: At left is the graph of the ceiling function and at right those of the fractional part function (again, in both figures one should ignore the vertical lines). Notice that the fractional part function is non-negative.



□

**5.E.58.** Examine the continuity of the function  $f(x) = (x - 1)^{-\operatorname{sgn}(x)}$ , at the points 0 and 1.

○

**5.E.59.** Given the points  $-\pi, 0, 1, 2, 3, \pi$ , determine whether the function

$$f(x) = \begin{cases} x, & x < 0; \\ 0, & 0 \leq x < 1; \\ x, & x = 1; \\ 0, & 1 < x < 2; \\ x, & 2 \leq x \leq 3; \\ \frac{1}{x-3}, & x > 3 \end{cases}$$

is continuous; left-continuous or right-continuous at these points.

○

**5.E.60.** Find all  $p \in \mathbb{R}$  for which the function

$$f(x) = \frac{\sin(6x)}{3x}, \quad x \in \mathbb{R}^* = \mathbb{R} \setminus \{0\}; \quad f(0) = p,$$

is continuous at the origin.

○

**5.E.61.** Choose a real number  $a$  so that the function

$$h(x) = \frac{x^4 - 1}{x - 1}, \quad x > 1; \quad h(x) = a, \quad x \leq 1,$$

is continuous on  $\mathbb{R}$ .

○

**5.E.62.** By defining the values at the points  $-1$  and  $1$ , extend the function

$$f(x) = (x^2 - 1) \sin \frac{2x - 1}{x^2 - 1}, \quad x \in \mathbb{R} \setminus \{\pm 1\},$$

so that the resulting function is continuous on the whole  $\mathbb{R}$ .

○

**5.E.63.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a function satisfying the relation

$$f^3(x) + 3f^2(-x)f(x) = 4x^2, \quad x \in \mathbb{R}. \quad (\#)$$

(a) Prove that  $f$  is even and next show that  $\lim_{x \rightarrow 0} f(x) = 0$ .

(b) Show that  $\lim_{x \rightarrow 0} \frac{\ln(f(x))}{\ln(\sqrt{x^3})} = 4/9$ .

**Solution.** (a) In the defining equation (#) replace  $x$  by  $-x$ . This gives  $f^3(-x) + 3f^2(x)f(-x) = 4x^2$  and by subtracting this relation from (#) (by parts), we get

$$f^3(x) - 3f^2(x)f(-x) + 3f(x)f^2(-x) - f^3(-x) = 4x^2 - 4x^2 = 0 \iff (f(x) - f(-x))^3 = 0, \quad x \in \mathbb{R},$$

where we used the identity  $(a - b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$ , with  $a, b \in \mathbb{R}$ . Therefore, we deduce that  $f(x) - f(-x) = 0$ , for all  $x \in \mathbb{R}$ , i.e.,  $f$  is even.

Now, because  $f$  is even the defining equation takes the form  $f^3(x) + 3f^2(x)f(x) = 4x^2$  or  $4f^3(x) = 4x^2$ , that is,  $f^3(x) = x^2$  for all  $x \in \mathbb{R}$ , from where we get  $f(x) = \sqrt[3]{x^2} = x^{\frac{2}{3}}$ ,  $x \in \mathbb{R}$ . As for the limit, recall that the function  $f(x) = x^{\frac{2}{3}}$  is everywhere continuous on  $\mathbb{R} = (-\infty, \infty)$ , see also the Problem 6.A.21 where the graph of  $f$  is presented. Therefore,  $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} x^{\frac{2}{3}} = f(0) = 0$ .

(b) The type of  $f$  is known by the first part, thus one can easily prove the claim, i.e.,

$$\lim_{x \rightarrow 0} \frac{\ln(f(x))}{\ln(\sqrt{x^3})} = \lim_{x \rightarrow 0} \frac{\ln(x)^{2/3}}{\ln(x)^{3/2}} = \lim_{x \rightarrow 0} \frac{(2/3)\ln(x)}{(3/2)\ln(x)} = \frac{4}{9} \cdot \lim_{x \rightarrow 0} \frac{\ln(x)}{\ln(x)} = \frac{4}{9}. \quad \square$$

**5.E.64.** Let  $f : \mathbb{R} \rightarrow (-\infty, 1)$  and  $g : \mathbb{R} \rightarrow (1, +\infty)$  be continuous functions satisfying  $f(a) = na$  and  $g(b) = nb$  for certain reals  $0 < a < b$  and some positive integer  $n$ . Prove the existence of some  $\xi \in (a, b)$  satisfying the equation

$$\xi = \frac{f(\xi)g(\xi)}{n}.$$

**Solution.** Observe that  $f(a) = na < 1$  since  $f(a) \in (-\infty, 1)$ , and hence  $a < \frac{1}{n}$ , while  $g(b) = nb > 1$  since  $g(b) \in (1, +\infty)$  and hence  $b > \frac{1}{n}$ . Thus, we can indeed form the interval  $[a, b]$  and consider the function

$$\ell(x) = f(x)g(x) - nx, \quad x \in [a, b].$$

Then we see that:

- $\ell(a) = f(a)g(a) - na = na(g(a) - 1) > 0$ , since  $g(a) > 1$  and so  $g(a) - 1 > 0$  and  $0 < a < 1/n$ .
  - $\ell(b) = f(b)g(b) - nb = nb(f(b) - 1) < 0$ , since  $f(b) < 1$  and hence  $f(b) - 1 < 0$  while  $0 < \frac{1}{n} < b$ .
- Thus,  $\ell(a)\ell(b) < 0$  and by Bolzano's theorem there exists some  $\xi \in (a, b)$  with  $\ell(\xi) = 0$ , that is,  $n\xi = f(\xi)g(\xi)$ .  $\square$

**5.E.65.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a periodic function with period  $T > 0$ , that is,  $f(x+T) = f(x)$  for all  $x \in \mathbb{R}$ . If  $\lim_{x \rightarrow +\infty} f(x)$  exists and is a real number  $\alpha$ , show that  $f(x) = \alpha$  for all  $x \in \mathbb{R}$ , i.e.,  $f$  is constant.<sup>14</sup>

**Solution.** By assumption  $f$  is periodic with period  $T$ , hence it suffices to prove the  $f$  is constant on  $[0, T]$ . Since  $\lim_{x \rightarrow +\infty} f(x) = \alpha \in \mathbb{R}$ , for every  $\varepsilon > 0$  there exist  $C > 0$  such that  $|f(x) - \alpha| < \varepsilon$ , for all  $x \geq C$ . Moreover, for all  $x \in [0, T]$  we can find some positive integer  $n$  such that  $x + nT \geq C$ . Thus we get  $|f(x) - \alpha| = |f(x + nT) - \alpha| < \varepsilon$ . This shows that  $|f(x) - \alpha| < \varepsilon$ , for all  $x \in [0, T]$  and  $\varepsilon > 0$ , and hence  $f(x) = \alpha$  for all  $x \in \mathbb{R}$ .  $\square$

**5.E.66.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function which is periodic with period  $T > 0$ . Show that there exists  $x_0 \in \mathbb{R}$  such that  $f(x_0 + \frac{T}{2}) = f(x_0)$ . Then use Sage to compute  $x_0$  for  $f(x) = \cos(x)$ .

**Solution.** It is sufficient to consider the function  $g(x) = f(x + \frac{T}{2}) - f(x)$  which is continuous on  $\mathbb{R}$  (since  $f$  is continuous) and satisfies

$$g(0) = f\left(\frac{T}{2}\right) - f(0), \quad g\left(\frac{T}{2}\right) = f(0) - f\left(\frac{T}{2}\right) = -g(0).$$

Recall that if a continuous function has values of opposite sign inside an interval, then it admits a root in that interval (Bolzano's theorem, see 5.2.19). In particular, the result follows by the intermediate value theorem applied to  $g$ , the latter restricted on  $[0, \frac{T}{2}]$ .

The  $\cos$  function  $f(x) = \cos(x)$  has period  $T = 2\pi$ , thus it suffices to focus on the interval  $(0, \frac{T}{2} = \pi)$ . The equation  $\cos(x_0 + \pi) = \cos(x_0)$  is equivalent to  $\cos(x_0) = 0$  (since  $\cos(x + \pi) = -\cos(x)$  for all  $x$ ). Hence obviously  $x_0 = \frac{\pi}{2} \in (0, \pi)$  (see also the table in 5.A.3).<sup>15</sup> In Sage just type `solve(cos(x + pi) == cos(x), x)`.  $\square$

<sup>14</sup>Later we will see that the condition  $\lim_{x \rightarrow +\infty} f(x) = \alpha \in \mathbb{R}$ , means that the line  $y = \alpha$  is a horizontal asymptote of  $y = f(x)$ .

<sup>15</sup>Recall that the equation  $\cos(x) = 0$  has general solution  $\frac{\pi}{2} + k\pi$ , with  $k \in \mathbb{N}$ .

**5.E.67.** Let  $f, g, h, k : \mathbb{R} \rightarrow \mathbb{R}$  continuous functions on  $\mathbb{R}$  satisfying the following conditions:

- $f(x) > 0$  and  $k(x) > 0$  for all  $x \in \mathbb{R}$ ;
- $f(a) = g(a)$  and  $h(b) = k(b)$  for some  $a \neq b \in \mathbb{R}$ ;
- $g(x) < f(x)$  for all  $x \in \mathbb{R} \setminus \{a\}$  and  $k(x) < h(x)$  for all  $x \in \mathbb{R} \setminus \{b\}$ .

Prove the existence of (at least) one  $x_0 \in (a, b)$  for which the vectors  $\vec{u} = (g(x_0), f(x_0))^T$  and  $\vec{v} = (k(x_0), h(x_0))^T$  (of  $\mathbb{R}^2$ ) are parallel.

**Solution.** The vectors  $\vec{u}$  and  $\vec{v}$  are parallel if and only if the determinant  $\begin{vmatrix} g(x_0) & f(x_0) \\ k(x_0) & h(x_0) \end{vmatrix}$  vanishes, i.e.,  $g(x_0)h(x_0) = k(x_0)f(x_0)$ . Consider the function  $\varphi(x) = g(x)h(x) - k(x)f(x)$ . This function is continuous on  $[a, b]$  and by the assumptions it follows that

$$\begin{aligned} \varphi(a)\varphi(b) &= (g(a)h(a) - k(a)f(a)) \cdot (g(b)h(b) - k(b)f(b)) \\ &= g(a)h(a)g(b)h(b) - g(a)h(a)k(b)f(b) - k(a)f(a)g(b)h(b) + k(a)f(a)k(b)f(b) \\ &= f(a)h(a)g(b)k(b) - f(a)h(a)k(b)f(b) - k(a)f(a)g(b)k(b) + k(a)f(a)k(b)f(b) \\ &= f(a)(h(a) - k(a)) \cdot k(b)(g(b) - f(b)) < 0. \end{aligned}$$

Thus, by the Bolzano's theorem there exists  $x_0 \in (a, b)$  with  $\varphi(x_0) = 0$  and the claim follows. □

**5.E.68.** Show that the equation  $x^{2024} + 3x - e^x - 1 = 0$  has at least a real solution in the open interval  $(0, 1)$ .

**Solution.** The function  $f(x) = x^{2024} + 3x - e^x - 1$  is sum of the polynomial  $x^{2024} + 3x - 1$  with the negative of the exponential function, hence by 5.2.17 it should be continuous on  $\mathbb{R}$  and hence also on the closed interval  $[0, 1]$ . Moreover, we have  $f(0) = -2 < 0$  and  $f(1) = 3 - e > 0$ . Therefore, the claim follows by the Bolzano's theorem. □

**5.E.69.** Let  $f$  be a real-valued continuous function defined on the closed interval  $[1, 4]$ .

- (a) Explain why the function  $g(x) = f(x) + e^{2x-1}$  with the same domain as  $f$  has a minimal and a maximal value there.
- (b) Prove the existence of some  $x_0 \in [1, 4]$  such that  $f(x) < f(x_0) + e^{2x_0}$ , for all  $x \in [1, 4]$ .

**Solution.** (a) The function  $g$  is continuous for all  $x \in [1, 4]$ , as the sum of two continuous function defined there, see the basic limit properties in 5.2.17. Thus  $g$  should take all the values between the maximal and the minimal one on the closed interval  $[1, 4]$ , by the Corollary in 5.2.19.

(b) Using the conclusion from the first part, we can find  $x_0 \in [1, 4]$  such that  $g(x) \leq g(x_0)$  for all such  $x$  (i.e., at  $x_0$  the function  $g$  attains its maximal value). This induces the result, i.e.,

$$f(x) \leq f(x_0) + e^{2x_0-1} - e^{2x-1} < f(x_0) + e^{2x_0-1} < f(x_0) + e^{2x_0}, \quad x \in [1, 4].$$

□

**5.E.70.** Let  $f : I \rightarrow \mathbb{R}$  be a continuous function defined on the interval  $I = [a, b]$ , where  $a, b \in \mathbb{R}$  with  $a < b$ . If  $f$  is injective, prove that  $f$  is strictly monotone. ○

**5.E.71.** Compute the following limits or explain why they do not exist:

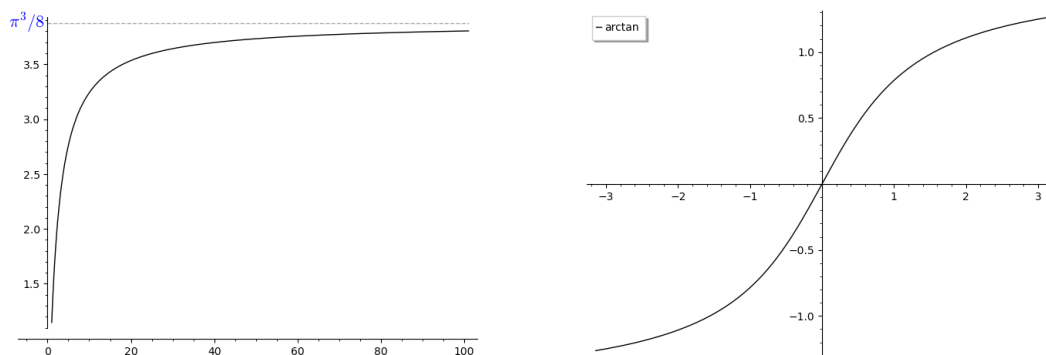
$$\lim_{x \rightarrow +\infty} (\arccos(1/(x+1)))^3, \quad \lim_{x \rightarrow -\infty} \arctan(1/x), \quad \lim_{x \rightarrow -\infty} \arctan(x^4), \quad \lim_{x \rightarrow -\infty} \arctan(\sin(x)).$$

Next, use Sage to verify your conclusions.

**Solution.** The function  $\arccos(x)$  is continuous on its domain  $[-1, 1]$ , while the function  $x^3$  is continuous everywhere. Moreover,  $\lim_{x \rightarrow +\infty} (1/(x+1)) = 0$ , and  $\arccos(0) = \pi/2$ . Thus,

$$\lim_{x \rightarrow +\infty} (\arccos(1/(x+1)))^3 = [\arccos(\lim_{x \rightarrow +\infty} 1/(x+1))]^3 = (\pi/2)^3.$$

The same conclusion occurs graphically, via the graph of the function  $f(x) = (\arccos(1/(1+x)))^3$ , see below at the left-hand-side for an illustration for  $x > 0$ :



Similarly, the function  $\arctan(x)$  is continuous and injective on its domain. Thus, according to 5.E.70 it should be strictly monotone, and indeed it is easy to verify that it is strictly increasing. This is also illustrated by the graph of  $\arctan(x)$ , as we see above at the right-hand-side. Now, according to the discussion in 5.2.16 these properties allow us to move the examined limit into the argument of such a function. Therefore, it is possible to write

$$\begin{aligned} \lim_{x \rightarrow -\infty} \arctan(1/x) &= \arctan\left(\lim_{x \rightarrow -\infty} (1/x)\right), \\ \lim_{x \rightarrow -\infty} \arctan(x^4) &= \arctan\left(\lim_{x \rightarrow -\infty} x^4\right), \\ \lim_{x \rightarrow -\infty} \arctan(\sin(x)) &= \arctan\left(\lim_{x \rightarrow -\infty} \sin(x)\right). \end{aligned}$$

However, we see that  $\lim_{x \rightarrow -\infty} (1/x) = 0$ ,  $\lim_{x \rightarrow -\infty} x^4 = +\infty$ , and  $\lim_{x \rightarrow -\infty} \sin(x)$  does not exist. Therefore, using the relations above we obtain  $\lim_{x \rightarrow -\infty} g(x) = \arctan(0) = 0$ ,  $\lim_{x \rightarrow -\infty} h(x) = \pi/2$ , while the last limit does not exist. Finally, to describe a confirmation via Sage just type

```
f(x)=(arccos(1/(1+x)))^3; print(lim(f(x), x=+oo))
g(x)=arctan(1/x); print(lim(g(x), x=-oo))
h(x)=arctan(x^4); print(lim(h(x), x=-oo))
k(x)=arctan(sin(x)); print(lim(k(x), x=-oo))
```

□

**5.E.72.** Show by an example that, in general, the intermediate value theorem fails for discontinuous functions.

**Solution.** For some positive real numbers  $\alpha, \delta$ , consider the function  $f : [-\delta, \delta] \rightarrow \mathbb{R}$  with

$$f(x) = \begin{cases} \alpha & \text{if } -\delta \leq x < 0, \\ -\alpha & \text{if } 0 \leq x \leq \delta. \end{cases}$$

Obviously, this function is discontinuous at  $x = 0$  since  $\lim_{x \rightarrow 0^-} f(x) = \alpha \neq \lim_{x \rightarrow 0^+} f(x) = -\alpha = f(0)$ . On the other hand we see that  $f(-\delta) = \alpha > 0$  and  $f(\delta) = -\alpha < 0$ , so that  $f(-\delta)f(\delta) < 0$ , but  $f(x) \neq 0$  for any  $x \in [-\delta, \delta]$ . □

Let us now present some further applications related to the intermediate value theorem. We first describe the 1-dimensional case of the so called Borsuk–Ulam theorem. We then study an application related to the temperature of antipodal points on the earth.

**5.E.73. 1-dimensional Borsuk–Ulam theorem.** Let  $S^1$  be the unit circle. Show that for any continuous function  $f : S^1 \rightarrow \mathbb{R}$  there exists some  $x \in S^1$  with  $f(-x) = f(x)$ .<sup>16</sup>

**Solution.** Consider the function  $g : S^1 \rightarrow \mathbb{R}$  defined by  $g(x) := f(x) - f(-x)$ . If  $g(x_0) = 0$  for some arbitrary  $x_0 \in S^1$ , we are done. Otherwise we may assume that  $g(x) > 0$  (and similarly is treated the case with  $g(x) < 0$ ). Then  $g(-x) = f(-x) - f(x) = -(f(x) - f(-x)) = -g(x) < 0$ . Thus, Bolzano’s theorem certifies the existence of some  $x_0 \in S^1$  with  $-x < x_0 < x$  satisfying  $g(x_0) = 0$ . This proves the claim.

We should mention that the result extends to  $n$  dimensions, as follows: Let

$$S^n = \left\{ (x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} : \sum_{k=1}^{n+1} x_k^2 = 1 \right\} \subset \mathbb{R}^{n+1}$$

<sup>16</sup>Prove that this is equivalent to say that any continuous odd function  $S^1 \rightarrow \mathbb{R}$  has a zero.

be the  $n$ -sphere. Then, for any continuous map  $f : S^n \rightarrow \mathbb{R}^n$  there exists  $x \in S^n$  such that  $f(-x) = f(x)$ .<sup>17</sup>  $\square$

**5.E.74. Earth's temperature and antipodal pints.** Prove that at any time there is a point on the globe where the temperature agrees with the temperature on the antipode, assuming that the temperature varies continuously.<sup>18</sup>

**Solution.** Consider the equator (or any meridian of the globe), and denote by  $T : [0, 2\pi] \rightarrow \mathbb{R}$  the temperature, so that  $T(x)$  is the temperature of a point  $x$  on this great circle. Notice that every great circle through any point also passes through its antipodal point, so there are infinitely many great circles through two antipodal points. Define the function  $f(x) = T(x) - T(x + \pi)$ , which is also continuous on  $[0, 2\pi]$ , since  $T$  is assumed to be continuous. If  $f(0) = 0$  we have found our antipodal points with equal temperature. Otherwise we may assume that  $f(0) \neq 0$ . We see that  $f(0) = T(0) - T(\pi)$ ,  $f(\pi) = T(\pi) - T(2\pi)$ , and since  $T$  is periodic with period  $2\pi$ , we finally get  $f(\pi) = T(\pi) - T(2\pi) = T(\pi) - T(0) = -f(0)$ . By Bolzano's theorem this implies that there exists some point  $x_0 \in (0, \pi)$  such that  $f(x_0) = 0$  (notice that  $T$  is also continuous on  $[0, \pi]$ ). Thus  $T(x_0) = T(x_0 + \pi)$ , which proves the claim.  $\square$

C) MATERIAL ON DERIVATIVES

*For convenience, we divide this paragraph into two subsections: one containing technical exercises for additional practice on derivatives, and the other focusing on applications of derivatives.*

C1) MATERIAL ON DERIVATIVES - PRACTICE

**5.E.75.** Show that the function  $f(x) = x|x|$  is differentiable for all  $x \in \mathbb{R}$ .

**Solution.** Based on the definition of the absolute value function, one deduces that  $f(x) = x^2$  for  $x > 0$  and  $f(x) = -x^2$  for  $x < 0$ , so for all  $x \in \mathbb{R} \setminus \{0\}$  the function  $f$  is differentiable (it is easy to see that  $(x^2)' = 2x$ , see also below). We need to check the differentiability of  $f$  at  $x_0 = 0$ , where  $f(0) = 0$ . We should compute the left and right derivatives,  $f'_-(0)$  and  $f'_+(0)$ , respectively. We see that

$$f'_-(0) = \lim_{x \rightarrow 0^-} \frac{f(x) - f(0)}{x - 0} = \lim_{x \rightarrow 0^-} \frac{-x^2 - 0}{x} = 0, \quad f'_+(0) = \lim_{x \rightarrow 0^+} \frac{f(x) - f(0)}{x - 0} = \lim_{x \rightarrow 0^+} \frac{x^2 - 0}{x} = 0,$$

that is, the left and right derivative exist and are equal. Hence,  $f$  is also differentiable at  $x_0 = 0$ , with  $f'(0) = 0$ .  $\square$

**5.E.76.** Compute the derivatives of the functions  $f(x) = (\tan(x))^x$  and  $g(x) = (\sin(x))^{\ln(x)}$ .

**Solution.** By the quotient rule and since  $\tan(x) = \frac{\sin(x)}{\cos(x)}$ , we easily get

$$\tan'(x) \equiv (\tan(x))' \equiv \frac{d}{dx} \tan(x) = \left( \frac{\sin(x)}{\cos(x)} \right)' = \frac{1}{\cos^2(x)}.$$

Moreover, we have  $f(x) = \tan(x) = e^{\ln(\tan(x))}$ , and hence we deduce that

$$f(x) = (e^{\ln(\tan(x))})^x = e^{x \ln(\tan(x))}.$$

Combining this with the relations  $(e^{h(x)})' = h'(x) \cdot e^{h(x)}$  and  $(\ln(h(x)))' = h'(x)/h(x)$ , we get

$$\begin{aligned} f'(x) &= \left( e^{x \ln(\tan(x))} \right)' = (x \ln(\tan(x)))' \cdot e^{x \ln(\tan(x))} = (x \ln(\tan(x)))' \cdot f(x) \\ &= \left( \ln(\tan(x)) + x \cdot \frac{\tan'(x)}{\tan(x)} \right) \cdot f(x) = \left( \ln(\tan(x)) + x \cdot \frac{1}{\tan(x)} \cdot \frac{1}{\cos^2(x)} \right) \cdot f(x) \\ &= \left( \ln(\tan(x)) + x \cdot \frac{(1 + \tan^2(x))}{\tan(x)} \right) \cdot f(x), \end{aligned}$$

where in the final relation one replaces  $1/\cos^2(x)$  by  $1 + \tan^2(x)$ . In Sage a verification is given as usual:

```
f(x)=tan(x)**x; df=diff(f, x); show(df)
```

In a similar vein, for the function  $g$  we get

$$g(x) = (\sin(x))^{\ln(x)} = (e^{\ln(\sin(x))})^{\ln(x)} = e^{\ln(x) \cdot \ln(\sin(x))}.$$

<sup>17</sup>Another famous result in topology with a very similar flavor is the so called "Brouwer fixed point theorem" which we will analyze in Chapter 9.

<sup>18</sup>Observe that the same result applies for the barometric pressure of antipodal points on the globe.

Thus,

$$\begin{aligned} g'(x) &= \left( (\sin(x))^{\ln(x)} \right)' = \left( e^{\ln(x) \cdot \ln(\sin(x))} \right)' = (\ln(x) \cdot \ln(\sin(x)))' \cdot g(x) \\ &= \left( \frac{1}{x} \cdot \ln(\sin(x)) + \ln(x) \cdot \cos(x) \cdot \frac{1}{\sin(x)} \right) \cdot g(x) = \left( \frac{\ln(\sin(x))}{x} + \ln(x) \cdot \cot(x) \right) \cdot g(x). \end{aligned}$$

For a verification in Sage type

```
g(x)=sin(x)**ln(x); show(g.derivative(x))
```

Observe that for your convenience the given solution in this task highlights two different, but familiar, methods in Sage that we can always use to compute derivatives. □

**5.E.77.** Differentiate the expression  $\frac{\sqrt[4]{x-1} \cdot (x+2)^3}{e^x(x+132)^2}$ , with  $x > 1$ . ○

**5.E.78.** Calculate the first derivative of the functions given below:

- (1)  $a(x) = (2 - x^2) \cos(x) + 2x \sin(x)$  with  $x \in \mathbb{R}$ ;
- (2)  $b(x) = \sin(\sin(x))$  with  $x \in \mathbb{R}$ ;
- (3)  $c(x) = \sin(\ln(x^3 + 2x))$  with  $x \in (0, +\infty)$ ;
- (4)  $d(x) = (1 + x - x^2)/(1 - x + x^2)$  with  $x \in \mathbb{R}$ ;
- (5)  $\varepsilon(x) = \sqrt{x \sqrt{x \sqrt{x}}}$  with  $x \in (0, +\infty)$ ;
- (6)  $f(x) = \sin(\sin(\sin(x)))$  with  $x \in \mathbb{R}$ ;
- (7)  $g(x) = \sqrt[3]{\sin(x)}$  with  $x \in \mathbb{R} \setminus \{n\pi : n \in \mathbb{Z}\}$ ;
- (8)  $h(x) = \sqrt[3]{\frac{1+x^3}{1-x^3}}$  with  $x \in \mathbb{R} \setminus \{\pm 1\}$ . ○

**5.E.79.** Recall that a function  $f$  is called *odd* if  $f(-x) = -f(x)$  for all  $x$  in its domain, and *even* if  $f(-x) = f(x)$  for all such  $x$ . Therefore, an odd function is symmetric about the origin, and an even function is symmetric with respect to the  $y$ -axis.

(a) Present a function defined on  $\mathbb{R}$  which is neither even, nor odd.

(b) Show that the derivative of an even (respectively odd) function is an odd (respectively even) function. ○

**5.E.80.** Extend the discussion of real polynomials of degree at most three, initiated in Problem 5.A.2, using derivatives. ○

So far we have discussed most of the transcendental functions, as the exponential, the logarithms, the trigonometric and their inverse. Another remarkable class consists of the so called “hyperbolic functions”, defined by



$$\sinh(x) = \frac{e^x - e^{-x}}{2}, \quad \cosh(x) = \frac{e^x + e^{-x}}{2}, \quad \tanh(x) = \frac{\sinh(x)}{\cosh(x)}, \quad \coth(x) = \frac{\cosh(x)}{\sinh(x)},$$

where  $x \in \mathbb{R}$  for the first three cases, while the domain of  $\coth$  is  $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ . Notice that

$$\cosh^2(x) - \sinh^2(x) = 1, \quad x \in \mathbb{R}.$$

The hyperbolic functions come at handy at times (especially in physics), and the next task is devoted to them, see also 5.4.12.

**5.E.81. Hyperbolic functions.** (a) Show that  $\sinh$  (respectively,  $\cosh$ ) is the odd (respectively, even) part of the exponential function. Moreover, prove that  $\tanh(-x) = -\tanh(x)$  for all  $x \in \mathbb{R}$  and  $\coth(x) = -\coth(x)$  for all  $x \in \mathbb{R}^*$ .

(b) Determine the derivatives of the hyperbolic functions on their domains.

(c) Confirm your answer in Sage and next sketch the graphs of  $\sinh$  and  $\cosh$  in the interval  $I = [-\pi, \pi]$ .

**Solution.** (a) Obviously, one has the relation

$$e^x = \frac{e^x - e^{-x}}{2} + \frac{e^x + e^{-x}}{2} = \sinh(x) + \cosh(x),$$

and is easy to see that  $\sinh(-x) = -\sinh(x)$ , while  $\cosh(-x) = \cosh(x)$  for all  $x \in \mathbb{R}$ . The rest two relations are now direct.

(b) This is really easy and relies on the basic rules of differentiation, see 5.3.4, in combination with the formula  $(e^x)' = e^x$ ,

for all  $x \in \mathbb{R}$ . Thus

$$(\sinh(x))' = \left( \frac{e^x - e^{-x}}{2} \right)' = \frac{1}{2} (e^x + e^{-x}) = \cosh(x),$$

$$(\cosh(x))' = \left( \frac{e^x + e^{-x}}{2} \right)' = \frac{1}{2} (e^x - e^{-x}) = \sinh(x),$$

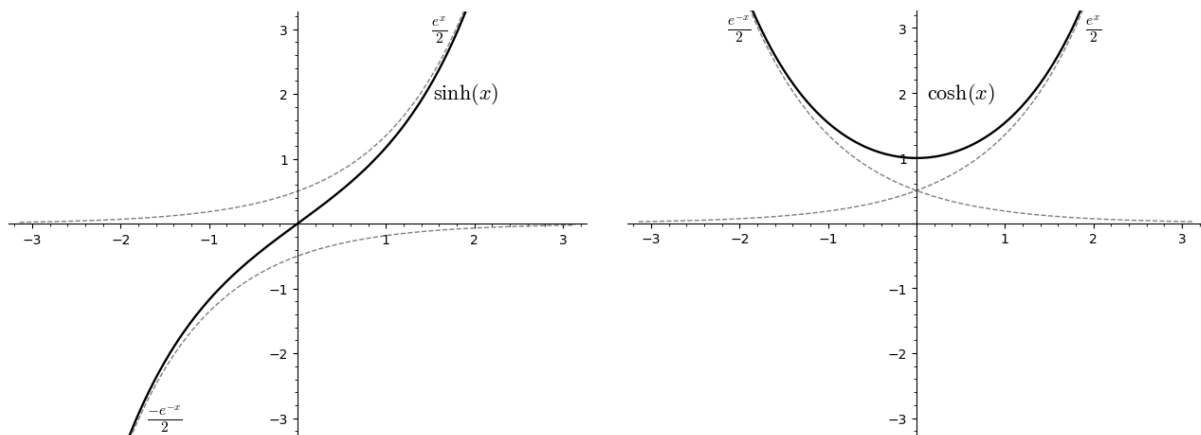
$$(\tanh(x))' = \left( \frac{\sinh(x)}{\cosh(x)} \right)' = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)} = \frac{1}{\cosh^2(x)} = 1 - \tanh^2(x),$$

$$(\coth(x))' = \left( \frac{\cosh(x)}{\sinh(x)} \right)' = \frac{\sinh^2(x) - \cosh^2(x)}{\sinh^2(x)} = -\frac{1}{\sinh^2(x)}.$$

(c) As one could expect in Sage the hyperbolic functions are built-in and correspond to the commands `sinh`, `cosh`, `tanh` and `coth`. Therefore, to confirm the previous computations in Sage one can combine these functions with either the command `diff`, or the command `derivative`, as follows

```
show(sinh(x).derivative()); show(cosh(x).derivative())
show(tanh(x).derivative()); show(coth(x).derivative())
```

Let us finally present the graphs of `sinh` and `cosh`, but leave for practice the corresponding coding in Sage.



□

**5.E.82.** Find the inverse of the hyperbolic sine function  $f(x) = \sinh(x)$ , determine its domain and its range, and next compute its first derivative. ○

**5.E.83.** Prove that the following function is differentiable at  $x = 0$ ,

$$f(x) = \begin{cases} x^2, & \text{if } x \in \mathbb{Q}, \\ 0, & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

**Solution.** Notice that  $f(0) = 0$ . Thus,  $\lim_{x \rightarrow 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x \rightarrow 0} \frac{f(x)}{x}$ . However, we see that  $0 \leq \left| \frac{f(x)}{x} \right| \leq \frac{|x^2|}{|x|} = |x|$ , and thus by the squeeze theorem we get  $\lim_{x \rightarrow 0} \frac{f(x)}{x} = 0$ . □

**5.E.84.** Show that the function  $f(x) = \sqrt{|x|}$  is *not* differentiable for all  $x \in \mathbb{R}$ . Which is the problematic point?

**Solution.** We see that  $f(x) = \sqrt{x}$  for  $x \geq 0$  and  $f(x) = \sqrt{-x}$  for  $x < 0$ . Thus for  $x \neq 0$  we may write

$$f'(x) = \begin{cases} \frac{1}{2\sqrt{x}}, & x > 0; \\ -\frac{1}{2\sqrt{-x}}, & x < 0. \end{cases}$$

However, at  $x_0 = 0$  the given function is not differentiable since for the left and right derivative one computes  $f'_-(0) = -\infty$  and  $f'_+(0) = +\infty$ , respectively (you should be able to verify the last claim by a formal computation, but also in Sage). □

**5.E.85.** Provide an example of a continuous function  $f : \mathbb{R} \rightarrow \mathbb{R}$  which is differentiable for all  $x \in \mathbb{R}$  but whose first derivative  $f'(x)$  is not continuous at  $x_0 = 0$ . Next use Sage to present the graph of your example.

**Solution.** An example is given by the piecewise function  $f : [-1, 1] \rightarrow \mathbb{R}$ , defined by

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right), & x \neq 0; \\ 0, & x = 0. \end{cases}$$

We know that  $\lim_{x \rightarrow 0} f(x) = 0$  and since  $f(0) = 0$ , the function  $f$  is continuous. A short computation shows that its first derivative is given by

$$f'(x) = \begin{cases} 2x \sin\left(\frac{1}{x}\right) - \cos\left(\frac{1}{x}\right), & x \neq 0; \\ 0, & x = 0. \end{cases}$$

Above, the first expression follows by the chain and product rule, while to verify that  $f'(0) = 0$ , notice that

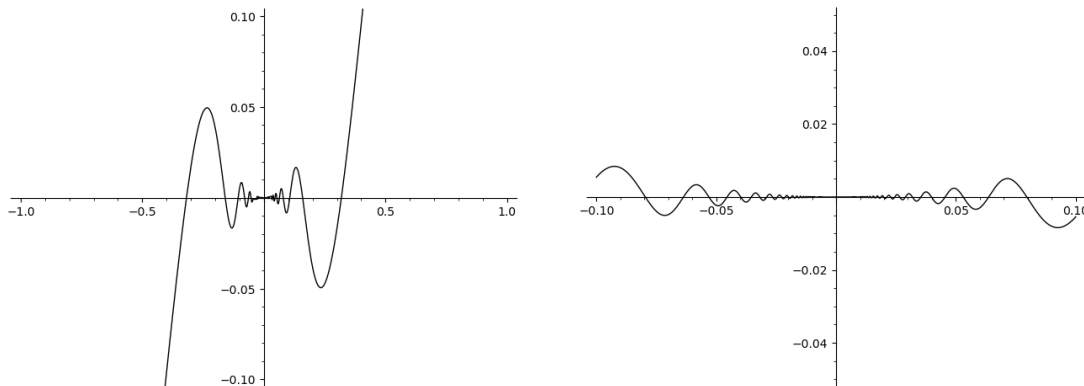
$$\lim_{x \rightarrow 0} \frac{f(x) - f(0)}{x} = \lim_{x \rightarrow 0} x \sin\left(\frac{1}{x}\right) = 0$$

(since  $0 \leq |x \sin(\frac{1}{x})| \leq |x|$ ). On the other hand, we see that the limit  $\lim_{x \rightarrow 0} (2x \sin(\frac{1}{x}) - \cos(\frac{1}{x}))$  does not exist, so  $f'(x)$  is not continuous at the origin. For another similar example see the description in 5.3.3.

Now, to present the graph of  $f$  one can apply the `piecewise` method described in 5.C.5. Thus we can proceed with the cell

```
f1(x)=(x^2)*sin(1/x); f2(x)=0
S=RealSet(x==0)
S1=RealSet.closed_open(-1,0)
S2=RealSet.open_closed(0,1)
S3=S1.union(S2)
F=piecewise([[S, f2(x)], [S3, f1(x)]])
p=plot(F(x), x, -1, 1, ymin=-0.1, ymax=0.1, color="black")
show(p)
```

In this block we appropriately used the command `RealSet` to produce the sets  $\{0\}$ ,  $[-1, 0)$ , and  $(0, 1]$ , and next the command `union` to introduce the set  $[-1, 0) \cup (0, 1]$ . Below, on the left is the figure produced by this block, while on the right we see a zoomed version of the graph of  $f$  near the origin. Can you realize the changes that we did in our code to get the zoomed version?



□

**5.E.86.** (1) Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(0) = 0$  and  $f(x) = x \arctan\left(\frac{1}{x}\right)$  for all  $x \in \mathbb{R} \setminus \{0\}$ . Determine whether the derivative of  $f$  exists at  $x_0 = 0$ .

(2) Consider the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(-1) = 0$  and  $f(x) = (x^2 - 1) \sin\left(\frac{1}{x+1}\right)$  for all  $x \in \mathbb{R} \setminus \{-1\}$ . Determine whether the derivative of  $f$  exists at  $x_0 = -1$ .

(3) Present an example of a function  $f : \mathbb{R} \rightarrow \mathbb{R}$  which is continuous on  $\mathbb{R}$  but does not have derivatives at the points  $x_1 = 5$  and  $x_2 = 9$ .

(4) Find functions  $f$  and  $g$  which are not differentiable anywhere, yet their composition  $f \circ g$  is differentiable everywhere on



$\mathbb{R}$ .

(5) For  $x > e$ , determine the sign of the derivative of the function  $f(x) = \arctan\left(\frac{\ln(x)}{-1 + \ln(x)}\right)$ . ○

**5.E.87.** If  $f(x) = 5x + 4 \sin(x)$ , then compute  $(f^{-1})'(4)$  by means of the formula of the derivative of the inverse function presented in 5.3.6. ○

**5.E.88.** (a) Let  $P(x)$  be a polynomial of order  $n \geq 2$ . Prove that  $P(x)$  is divided by  $(x - \rho)^2$  with quotient  $\pi(x)$ , i.e.,  $P(x) = (x - \rho)^2\pi(x)$ , if and only if  $P(\rho) = P'(\rho) = 0$ .

(b) Show that  $(x - 1)^2$  is a factor of the polynomial  $P(x) = nx^{n+1} - x^{n-1} - (n^2 + 1)x + n^2 - n + 2$ , with  $n \geq 2$ .

**Solution.** (a) If  $P(x) = (x - \rho)^2\pi(x)$  then by differentiation we get  $P'(x) = 2(x - \rho)\pi(x) + (x - \rho)^2\pi'(x)$ . Thus  $P'(\rho) = 0 = P(\rho)$ . Conversely, assume that  $P(\rho) = 0 = P'(\rho)$ . From the first relation we get  $P(x) = (x - \rho)\nu(x)$  for some polynomial  $\nu(x)$ , and hence  $P'(x) = \nu(x) + (x - \rho)\nu'(x)$ . But then the second relation can be equivalently written as  $\nu(\rho) + (\rho - \rho)\nu'(\rho) = 0$ , that is  $\nu(\rho) = 0$ . Therefore  $\rho$  is a root of  $\nu(x)$  and hence  $\nu(x) = (x - \rho)\pi(x)$  for some polynomial  $\pi(x)$ . It follows that  $P(x) = (x - \rho)^2\pi(x)$ , as it is required.

(b) We see that  $P(1) = n - 1 - n^2 - 1 + n^2 - n + 2 = 0$ . Moreover, the first derivative of  $P$  is given by  $P'(x) = n(n + 1)x^n - (n - 1)x^{n-2} - (n^2 + 1)$ , and thus  $P'(1) = n(n + 1) - (n - 1) - n^2 - 1 = n^2 + n - n + 1 - n^2 - 1 = 0$ . The claim now follows from the statement in (a). □

**5.E.89.** Let  $P$  be a cubic polynomial satisfying the conditions  $P(0) = 1$ ,  $P'(0) = 1$ ,  $P(1) = 2a + 2$ ,  $P'(1) = 5a + 1$ . Find the values of the parameter  $a \in \mathbb{R}$  for which the polynomial  $P$  is strictly monotonic on the whole real line. ○

**5.E.90.** Suppose that  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  are two functions differentiable at  $a \in \mathbb{R}$ . Prove that  $\lim_{x \rightarrow a} \frac{x^2 f(a) - a^2 f(x)}{x - a} = 2af(a) - a^2 f'(a)$ , and more in general

$$\lim_{x \rightarrow a} \frac{g(x)f(a) - g(a)f(x)}{x - a} = g'(a)f(a) - f'(a)g(a).$$

**Solution.** For the first limit we have

$$\begin{aligned} \lim_{x \rightarrow a} \frac{x^2 f(a) - a^2 f(x)}{x - a} &= \lim_{x \rightarrow a} \frac{x^2 f(a) - a^2 f(x) - a^2 f(a) + a^2 f(a)}{x - a} = \lim_{x \rightarrow a} \frac{(x^2 - a^2)f(a) - a^2(f(x) - f(a))}{x - a} \\ &= f(a) \lim_{x \rightarrow a} \frac{x^2 - a^2}{x - a} - a^2 \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} = f(a) \lim_{x \rightarrow a} (x + a) - a^2 f'(a) \\ &= 2af(a) - a^2 f'(a). \end{aligned}$$

For the general case we apply the same trick:

$$\begin{aligned} \lim_{x \rightarrow a} \frac{g(x)f(a) - g(a)f(x)}{x - a} &= \lim_{x \rightarrow a} \frac{g(x)f(a) - g(a)f(x) - g(a)f(a) + g(a)f(a)}{x - a} \\ &= \lim_{x \rightarrow a} \frac{(g(x) - g(a))f(a) - g(a)(f(x) - f(a))}{x - a} \\ &= f(a) \lim_{x \rightarrow a} \frac{g(x) - g(a)}{x - a} - g(a) \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} = f(a)g'(a) - g(a)f'(a). \end{aligned}$$

□

**5.E.91.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a differentiable function at 0, satisfying  $f(x)f(y) \neq 1$  for all  $x, y \in \mathbb{R}$  and

$$f(x + y)(1 - f(x)f(y)) = f(x) + f(y), \quad x, y \in \mathbb{R}.$$

Show that  $f$  is differentiable on the whole real line. ○

**5.E.92. Interactive tangent lines via Sage.** For the polynomial  $P(x) = x^4 - 2x$  use Sage to produce an interactive graph for the tangent line of  $P$  at a random point  $x_0$ , lying in the interval  $[-2, 2]$ . (Hints: Combine the routine constructed in 5.C.14 with the method presented in ?? for creating an interactive plot).

**Solution.** Let us use the routing introduced in 5.C.14 for the construction of the tangent line and combine this with the commands `@interact` and `slider`, which are responsible for the creation and the control of the interactive environment, respectively. In particular, the command `@interact` appears always in a separate line inside the syntax, usually before the `def` command that our subroutine starts. On the other hand, the `slider` command converts the `x0` input from a numerical input into a slider. The coding goes as follows

```
P(x)=x^4-2*x; P1(x)=diff(P(x), x)
@interact
```

```
def Tangentat(x_0 = slider(-2, 2, 0.1, -1.5, label="x-coordinate")) :
    y_0=P(x_0)
    m=P1(x_0)
    c=y_0 - m*x_0
    l(x)=m*x+c
    Q = plot(P(x), -2, 2, color="blue", ymin=-2, ymax=20)
    Q+=plot(l(x), -2, 2, color="black", ymin=-2, ymax=20)
    Q+=point((x_0, y_0), color="black", size=50)
    Q.show()
    show("x_0=" , x_0)
    show("tang. line: l(x) = (" , m, ") * x + (" , c, ")")
    return
```

Recall inside the command `slider` the first two parameters determine the interval where that variable  $x_0$  is defined, in our case  $-2 \leq x \leq 2$ . The third parameter determines the granularity of the slider, and our choice represents a movement with step 0.1. Notice in CoCalc.com one should call the subroutine by typing `Tangentat()` in the very end of the block presented above (without specifying  $x_0$  inside the parenthesis). However, in <https://sagecell.sagemath.org> this is not necessary and one can simply move the slider around. Test yourself Sage's output.  $\square$

**5.E.93.** Consider the functions  $f(x) = 2 + \ln(x - 1)$  and  $g(x) = 2 - \ln(x - 1)$ , with  $x \in A = (1, +\infty)$ .

(a) Use Sage to find the common points of the graphs of  $f, g$ , respectively.

(b) Deduce about the monotonicity of  $f, g$ , and show that  $x = 1$  is an asymptote without slope (vertical asymptote) both of  $C_f$  and  $C_g$ . Moreover, deduce that  $f(A) = g(A) = \mathbb{R}$ .

(c) Show that the tangent lines of  $f, g$  at points with the same  $y$ -coordinate are perpendicular.

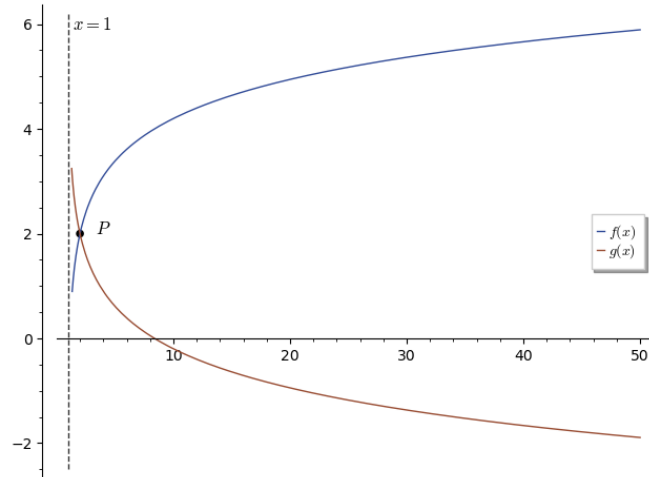
**Solution.** (a) This can be done by solving the equation  $f(x) = g(x)$ , which is equivalent to  $\ln(x - 1) = 0$  and has  $x = 2$  as a unique solution. There we see that  $f(2) = g(2) = 2$ , hence the common point of  $C_f$  and  $C_g$  is the point  $P = [2, 2]$ . A solution via Sage occurs by the cell

```
f(x)=2+ln(x-1); g(x)=2-ln(x-1); solve(f(x)==g(x), x)
```

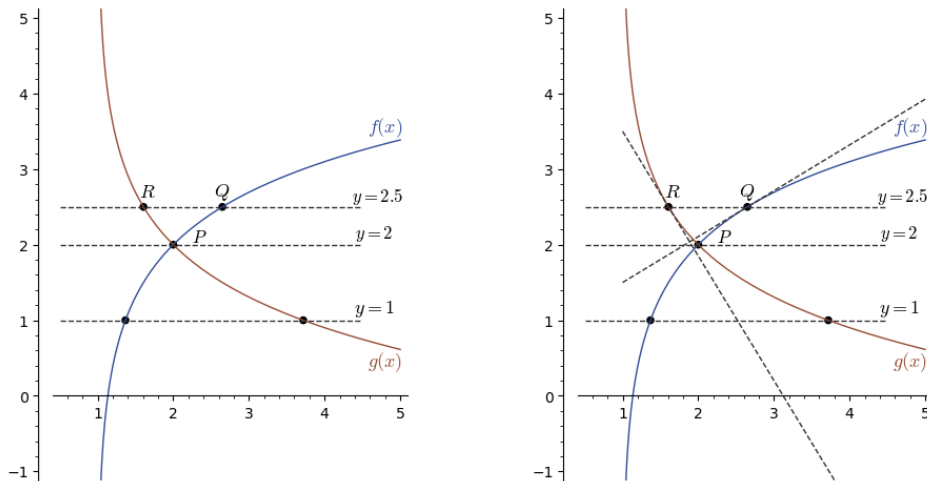
(b) We see that  $f'(x) = \frac{1}{x-1} > 0$  and  $g'(x) = -\frac{1}{x-1} < 0$  for all reals  $x > 1$ . Thus  $f$  is strictly increasing on  $A$ , while  $g$  is strictly decreasing on  $A$ . Moreover, you can compute  $a = \lim_{x \rightarrow 1^+} f(x) = -\infty, b = \lim_{x \rightarrow \infty} f(x) = +\infty, c = \lim_{x \rightarrow 1^+} g(x) = +\infty$  and  $d = \lim_{x \rightarrow \infty} g(x) = -\infty$ . These limits can be confirmed in Sage, as usual:

```
f(x)=2+ln(x-1); g(x)=2-ln(x-1);
a=lim(f(x), x=1, dir="right"); b=lim(f(x), x=oo);
c=lim(g(x), x=1, dir="right"); d=lim(g(x), x=oo)
show(a, ", ", b, ", ", c, ", ", d)
```

Both functions  $f, g$  are continuous, hence by the limits above we deduce that  $f(A) = \mathbb{R}$  and  $g(A) = \mathbb{R}$ . Moreover, the limits  $a, c$  presented above show that the graphs of  $f, g$ , respectively, tend toward  $\mp\infty$  as the inputs approach 1, and hence the line  $x = 1$  is a vertical asymptote for both  $C_f, C_g$  (we will discuss more details on asymptotes in 6.1.8). To illustrate the behaviour of  $f, g$ , we used Sage to sketch the graphs  $C_f, C_g$ , for  $1 \leq x \leq 50$ , which we present together here:



(c) To get points on  $C_f, C_g$  with the same  $y$ -coordinate it is sufficient to draw horizontal lines  $y = c, c \in \mathbb{R}$ . This happens due to the type of monotonicity of  $f, g$ . Let  $Q, R$  be the corresponding intersection points of the line  $y = c$  with  $C_f$  and  $C_g$ , respectively. We may assume that  $Q = [a, f(a)] = [a, c]$  and  $R = [b, g(b)] = [b, c]$ . Notice for  $a = 2 = b$  we have  $P = Q = R$ . The getter a better view around the intersection point of  $C_f$  and  $C_g$ , see the figure below (at the l.h.s). In this figure, to become more precise, we fix the points  $Q, R$  to be elements of the line  $y = 2.5$  (hence one can specify  $Q, R$  explicitly, as we did).



From the mathematical point of view it is not necessary to fix  $c$ . At the the points  $Q = [a, c] \in C_f$  and  $R = [b, c] \in C_g$  we have  $f'(a) = \frac{1}{1-a}$  and  $g'(b) = -\frac{1}{1-b}$ , and to prove the claim it is sufficient to show hat  $f'(a)g'(b) = -1$ . Because  $f(a) = c = g(b)$  we get  $a - 1 = e^{c-2}, b - 1 = e^{2-c}$  and the result follows (for an illustration see the picture at the r.h.s, which includes the tangent lines of  $C_f, C_g$ , at the points  $Q, R$ , respectively), i.e.,

$$f'(a)g'(b) = -\frac{1}{(1-a)(1-b)} = -\frac{1}{e^{c-2} e^{2-c}} = -\frac{1}{e^0} = -1.$$

For your convenience, let us finally present the code used to construct the figure including the tangent lines.

```
f(x)=2+ln(x-1); g(x)=2-ln(x-1);
p=plot(f(x), x, 1, 5, rgbcolor=(0.2,0.3,0.6))
p+=plot(g(x), x, 1, 5, rgbcolor=(0.6,0.3,0.2))
p+=point((2, 2), color="black", size=30)
p+=text(r"$f(x)$", (4.8, 3.55), rgbcolor=(0.2,0.3,0.6), fontsize=13)
p+=text(r"$g(x)$", (4.8, 0.45), rgbcolor=(0.6,0.3,0.2), fontsize=13)
p+=text(r"$P$", (2.35, 2.1), color="black", fontsize=12)
p+=line([(0.5, 1), (4.5, 1)], rgbcolor=(0.2,0.2,0.2), linestyle="--")
p+=line([(0.5, 2.5), (4.5, 2.5)], rgbcolor=(0.2,0.2,0.2), linestyle="--")
p+=line([(0.5, 2), (4.5, 2)], rgbcolor=(0.2,0.2,0.2), linestyle="--")
```

```

p+=point((e^(1/2) + 1, f(e^(1/2) + 1)), color="black", size=30)
p+=text(r"$Q$", (e^(1/2) + 1, f(e^(1/2) + 1)+0.2), color="black", fontsize=13)
p+=point(((e^(1/2) + 1)*e^(-1/2), g((e^(1/2) + 1)*e^(-1/2))), color="black", size=30)
p+=text(r"$R$", ((e^(1/2) + 1)*e^(-1/2)+0.05, g((e^(1/2) + 1)*e^(-1/2))+0.2),
color="black", fontsize=13)
p+=text(r"$y=2.5$", (4.7, 2.65), color="black", fontsize=12)
p+=text(r"$y=2$", (4.65, 2.15), color="black", fontsize=12)
p+=point(((e + 1)*e^(-1), f((e + 1)*e^(-1))), color="black", size=30)
p+=point((e + 1, g(e + 1)), color="black", size=30)
p+=text(r"$y=1$", (4.65, 1.15), color="black", fontsize=13)
xQ=e^(1/2) + 1
tangf=f(xQ)+diff(f, x)(x=xQ)*(x-xQ)
p+=plot(tangf, x, 1, 5, rgbcolor=(0.2,0.2,0.2), linestyle="--")
xR=(e^(1/2) + 1)*e^(-1/2)
tangg=g(xR)+diff(g, x)(x=xR)*(x-xR)
p+=plot(tangg, x, 1, 5, rgbcolor=(0.2,0.2,0.2), linestyle="--")
p.show(ymin=-1, ymax=5, ticks=[1, 1], aspect_ratio=1)
    
```

□

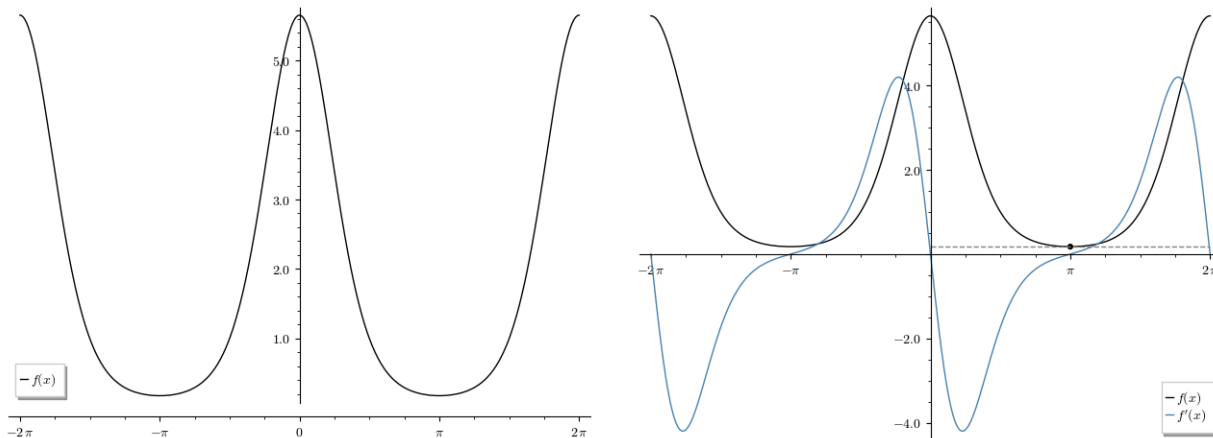
**5.E.94.** Consider the function  $f(x) = e^{\sqrt{\alpha} \cos(x)}$ , where  $\alpha > 0$  is some fixed constant. Provide a proof of the following facts::

- $f$  is continuous, even, and periodic with period independent of  $\alpha$ . Sketch the graph of  $f$  for  $\alpha = 3$  and  $x \in [-2\pi, 2\pi]$ ;
- $f$  is differentiable for all  $x \in \mathbb{R}$ . Find the formula of  $f'(x)$ ;
- $f$  is (strictly) decreasing in  $(0, \pi)$  and (strictly) increasing in  $(\pi, 2\pi)$ , in particular, the point  $x_0 = \pi$  provides a local minimum of  $f$ .

**Solution.** (a) The domain of  $f$  is the whole real line  $\mathbb{R}$ . Since  $f$  is the composition of continuous functions, it is also continuous. It is even since  $\cos(-x) = \cos(x)$  which implies that  $f(-x) = f(x)$  for all  $x \in \mathbb{R}$ . Suppose that  $T \in \mathbb{R} \setminus \{0\}$  satisfies  $f(x+T) = f(x)$ , for all  $x \in \mathbb{R}$ . Since the exponential map is an injection this gives  $\sqrt{\alpha} \cos(x+T) = \sqrt{\alpha} \cos(x)$ , that is,  $\cos(x+T) = \cos(x)$  for all  $x \in \mathbb{R}$ . Thus

$$x+T = 2k\pi + x, \quad \text{or} \quad x+T = 2k\pi - x,$$

which implies that  $T = 2k\pi$  or  $T = 2k\pi - 2x$  for all  $x \in \mathbb{R}$  and some  $k \in \mathbb{Z}$ . This means that  $f$  is periodic with period  $2k\pi$ , so for  $k = 1$  we obtain  $T = 2\pi$  which is the smallest (positive) period of  $f$ . A visual verification of this fact occurs by sketching the graph of  $f$ , for some fixed  $\alpha$ . Below we present the case for  $\alpha = 3$ .



- The function  $f$  is differentiable on  $\mathbb{R}$  as the composition of differentiable functions. By the chain rule we see that

$$f'(x) = (\sqrt{\alpha} \cos(x))' e^{\sqrt{\alpha} \cos(x)} = -\sqrt{\alpha} \sin(x) \cdot e^{\sqrt{\alpha} \cos(x)}, \quad x \in \mathbb{R}.$$

Or we can compute this in Sage, via the cell

```
a=var("a"); f(x)=e**(sqrt(a)*cos(x)); show(diff(f, x))
```

- In  $[0, 2\pi]$  the equation  $f'(x) = 0$  has three solutions, namely the points  $x = 0$ ,  $x = \pi$  and  $x = 2\pi$ , see also the figure in the r.h.s. where we have included the graph of  $f'(x)$ . Moreover we see that  $f'(x) < 0$  for all  $x \in (0, \pi)$  and  $f'(x) > 0$  for all

$x \in (\pi, 2\pi)$ . It follows that  $f(x)$  is decreasing in the first interval, and increasing in the second one. Hence  $x_0 = \pi$  is a local minimum of  $f$  giving the minimal value  $f(\pi) = e^{-\sqrt{\alpha}}$  (the other two stationary points are local maxima of  $f$  in  $[0, 2\pi]$ ).  $\square$

- 5.E.95.** (a) Determine all local extrema of the function  $f(x) = x \ln^2(x)$  with domain  $(0, +\infty)$ ;  
 (b) Determine the maximum value of the function  $f(x) = e^{-x} \cdot \sqrt[3]{3x}$ , with  $x \in \mathbb{R}$ .  
 (c) Find the absolute extrema of the polynomial  $p(x) = x^3 - 3x + 2$  on the interval  $[-3, 2]$ .  
 (d) Is there any  $\alpha \in \mathbb{R}$  such that the function  $h(x) = \alpha x + \sin(x)$ ,  $x \in [0, 2\pi]$  has a global minimum at  $x_0 = 5\pi/4$ ?  $\circ$

**5.E.96.** Show that the function  $f(x) = e^{-x}$  is strictly decreasing for any  $x \in \mathbb{R}$ .  $\circ$

**5.E.97.** Find all the solutions of the equation  $x^{2025} + 2024x = 2025 - \ln(x)$ .  $\circ$

**5.E.98.** Let  $I = [a, b] \subset \mathbb{R}$  be a closed subset of  $\mathbb{R}$  and let  $f : I \rightarrow \mathbb{R}$  be a continuous function. Suppose that  $f$  is differentiable on the open interval  $(a, b)$  and satisfies  $f(a) = f(b) = 0$ . Show that for given  $\kappa \in \mathbb{R}$  there exists  $x_0 \in (a, b)$  such that  $\kappa f(x_0) + f'(x_0) = 0$ .

**Solution.** For fixed  $\kappa \in \mathbb{R}$  consider the function  $g(x) = e^{\kappa x} f(x)$ , with  $x \in [a, b]$ . By assumption  $f$  is continuous on  $[a, b]$ , differentiable on  $(a, b)$ , and the same is  $g$ . We see that

$$g'(x) = \kappa e^{\kappa x} f(x) + e^{\kappa x} f'(x) = e^{\kappa x} (\kappa f(x) + f'(x)), \quad x \in (a, b). \quad (*)$$

Moreover,  $g(a) = 0 = g(b)$  and by Rolle's theorem (see 5.3.8) one can find some  $x_0 \in (a, b)$  such that  $g'(x_0) = 0$ . Then the result is a simple consequence of (\*).  $\square$

**5.E.99.** Let  $f : [0, 4] \rightarrow \mathbb{R}$  be a differentiable function with  $f(0) = f(4) = 0$ , and set  $g(x) := \frac{x(4-x)}{4} f(2)$ , for any  $x \in [0, 4]$ . Show that there exist  $x_1, x_2 \in (0, 4)$  with  $x_1 \neq x_2$  satisfying  $f'(x_1) = g'(x_1)$  and  $f'(x_2) = g'(x_2)$ .

**Solution.** The functions  $f, g$  are both differentiable on  $[0, 4]$ , so the same is their difference  $h(x) := f(x) - g(x) = f(x) - \frac{x(4-x)}{4} f(2)$ . Moreover, we see that  $h(0) = f(0) - g(0) = 0$ ,  $h(4) = f(4) - g(4) = 0$  and  $h(2) = f(2) - \frac{2 \cdot 2}{4} f(2) = 0$ . Thus,  $h$  satisfies the Rolle's theorem conditions on the intervals  $[0, 2]$  and  $[2, 4]$ . This means that there exist  $x_1 \in (0, 2)$  such that  $h'(x_1) = 0$ , which is equivalent to  $f'(x_1) = g'(x_1)$  and  $x_2 \in (2, 4)$  such that  $h'(x_2) = 0$ , that is,  $f'(x_2) = g'(x_2)$ .  $\square$

**5.E.100.** Suppose that  $f : I \rightarrow \mathbb{R}$  is a continuous function on  $I = [a, b]$ , differentiable on  $(a, b)$  and such that  $\lim_{x \rightarrow a} f'(x) = \kappa \in \mathbb{R}$ . Show that  $f$  is also differentiable at  $x = a$  with  $f'(a) = \kappa$ .

**Solution.** We can apply the mean value theorem on  $f$  on the interval  $(a, a+h)$  for some small enough  $h$ . Hence there exists  $c_h \in (a, a+h)$  such that

$$f'(c_h) = \frac{f(a+h) - f(a)}{h}. \quad (*)$$

Moreover, we see that  $c_h \rightarrow a$  as  $h \rightarrow 0$  and hence

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \stackrel{(*)}{=} \lim_{h \rightarrow 0} f'(c_h) = \lim_{x \rightarrow a} f'(x) = \kappa. \quad \square$$

**5.E.101.** Show that the polynomial  $P(x) = x^5 - x^4 + 2x^3 - x^2 + x + 1$  has exactly one real root. Next use Sage to approximate this root.

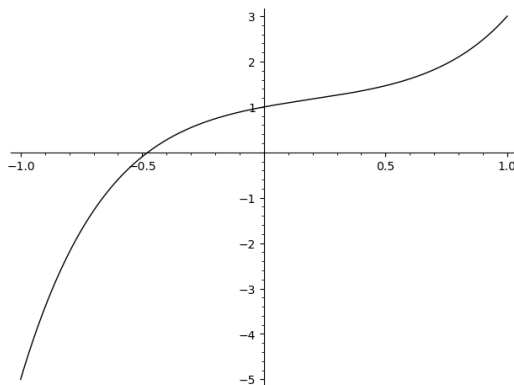
**Solution.** It is easy to prove that any odd degree polynomial has at least one real root. For instance, the polynomial at hand has degree 5, hence by the fundamental theorem of algebra (see 12.4.20),  $P$  has to have five roots over  $\mathbb{C}$ . Moreover, the complex roots with real coefficients come in pairs of conjugated numbers. Therefore,  $P$  has to have at least one real root, say  $x_0$ .

Suppose now that there exists a second real root  $x'_0$ . Then, according to the Rolle's theorem there must be some  $c \in (x_0, x'_0)$ , such that  $P'(c) = 0$ . But  $P'(x) = 5x^4 - 4x^3 + 6x^2 - 2x + 1 = 2x^2(x-1)^2 + 3x^4 + 3x^2 + (x-1)^2 > 0$ , for all  $x \in \mathbb{R}$ , a contradiction. Therefore,  $P$  admits exactly one real root. In other words,  $P$  is strictly increasing and thus its graph should cross the  $x$ -axis just once.

From the graph of  $P$  given below we see that the unique root of  $P$  should be near to  $-0.5$ . In Sage to obtain this root we can use the `find_root` function, as follows:

```
find_root(x^5-x^4+2*x^3-x^2+x+1, -1, 1)
```

According to Sage's answer, the unique root of  $P$  is given by  $x_0 \approx -0.47748$ .



□

**5.E.102.** Let  $f : \mathbb{R} \rightarrow (0, +\infty)$  be a continuously differentiable function with  $f(0) \neq 0$ . Prove that there exists  $\xi \in (0, 1)$  satisfying the relation  $e^{f'(\xi)} f(0)^{f(\xi)} = f(1)^{f(\xi)}$ .

**Solution.** Let us rewrite the equation as  $e^{f'(\xi)} = \left(\frac{f(1)}{f(0)}\right)^{f(\xi)}$ . By applying the logarithm in both sides we get

$$\ln\left(e^{f'(\xi)}\right) = f'(\xi) = \ln\left(\frac{f(1)}{f(0)}\right)^{f(\xi)} \iff f'(\xi) = f(\xi) \ln\left(\frac{f(1)}{f(0)}\right) \iff \frac{f'(\xi)}{f(\xi)} = \ln(f(1)) - \ln(f(0)).$$

It is now obvious that the existence of such a  $\xi$  is guaranteed by applying the Lagrange mean value theorem to the function  $f(x) = \ln(f(x))$ , which is continuous on  $[0, 1]$  and differentiable at all points inside this interval. □

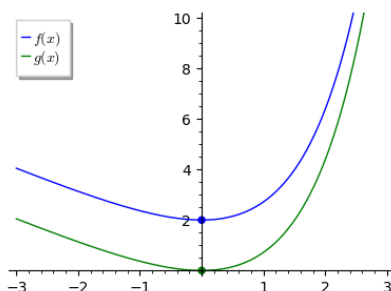
**5.E.103.** Show that any real  $x \geq 0$  satisfies  $x \geq \ln(x + 1)$ . ○

**5.E.104.** Let  $\alpha, \beta$  be reals satisfying  $0 < \alpha < \beta$ . Show that  $\alpha x^\beta - \beta x^\alpha > \alpha - \beta$ , for all  $x > 1$ . ○

**5.E.105.** Show that all  $x \in \mathbb{R}$  satisfy  $e^x - x \geq 1$ .

**Solution.** Consider the function  $f(x) = e^x - x + 1$ . This is differentiable in  $\mathbb{R}$  with  $f'(x) = e^x - 1$ . We see that the equation  $f'(x) = 0$  has a unique solution given by  $x = 0$ . In particular, having in mind the graph of  $e^x$  we see that  $e^x < 1$  for  $x < 0$ , so  $f'(x) < 0$  for  $x \in (-\infty, 0)$  and  $e^x > 1$  for  $x > 0$  which means that  $f'(x) > 0$  for  $x \in (0, +\infty)$ . Hence  $f$  is strictly decreasing in the first interval and strictly increasing in the second one, see the figure below. Thus, for  $x > 0$  we have  $f(x) > f(0)$  and so  $e^x - x + 1 > 2$ , that is  $e^x - x > 1$ , while for  $x < 0$  we have  $f(x) > f(0)$  and so again  $e^x - x + 1 > 2$ . So for  $x \neq 0$  we get  $e^x - x > 1$ . Since for  $x = 0$  this holds as an equality we have finally proved  $e^x - x \geq 1$  for all  $x \in \mathbb{R}$ .

**Remark:** The graphs of  $f, g$  sketched below indicate that one may use the function  $g(x) = e^x - x - 1$ , instead of  $f$ , and apply the same procedure. Try to verify this claim.



□

**5.E.106.** For  $x > y > 0$  prove the inequality

$$\frac{x + y}{2} > \frac{x - y}{\ln(x) - \ln(y)}.$$

**Solution.** Rewrite the given inequality as  $\frac{\ln(x) - \ln(y)}{x - y} > \frac{2}{x + y}$  and set  $t = x - y > 0$ . Then we get

$$\frac{\ln(t + y) - \ln(y)}{t} > \frac{2}{t + 2y} \iff \frac{\ln\left(\frac{t+y}{y}\right)}{t} > \frac{2}{t + 2y} \iff \ln\left(1 + \frac{t}{y}\right) > \frac{2t}{t + 2y} \iff \ln\left(1 + \frac{t}{y}\right) > \frac{2\frac{t}{y}}{2 + \frac{t}{y}}.$$



Putting now  $\psi = \frac{t}{y}$  we have  $\psi > 0$  and the initial inequality can be equivalently transformed to

$$\ln(1 + \psi) > \frac{2\psi}{2 + \psi}, \quad \psi > 0. \quad (b)$$

Thus now you may consider the function  $f(\psi) = \ln(1 + \psi) - \frac{2\psi}{2 + \psi}$ , with  $\psi > -1$ . Its first derivative is given by

$$f'(\psi) = \frac{1}{\psi + 1} - \frac{2(2 + \psi) - 2\psi}{(2 + \psi)^2} = \frac{\psi^2}{(\psi + 1)(2 + \psi)^2} > 0,$$

that is,  $f'(\psi) > 0$  for all  $\psi > -1$ . Hence for example we should have  $f(\psi) > f(0) = 0$ , which gives us the inequality presented in (b).  $\square$

**5.E.107.** If  $p, q \in \mathbb{R}$  are such that  $0 < p < q$  show that

$$\ln\left(\frac{p+q}{2}\right) < \frac{p}{p+q} \ln(p) + \frac{q}{p+q} \ln(q). \quad (\#)$$

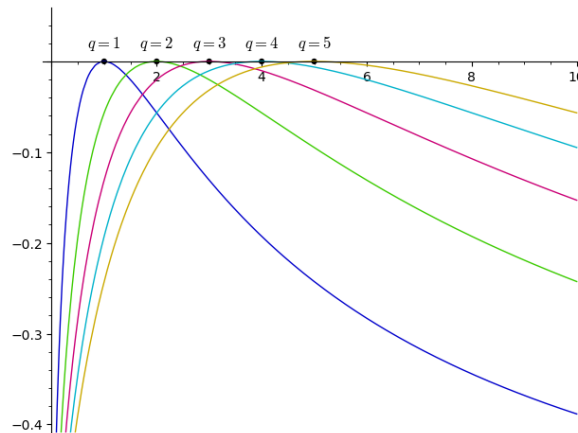
**Solution.** Consider the 1-parameter family  $f_q(x) = \ln\left(\frac{x+q}{2}\right) - \frac{x}{x+q} \ln(x) - \frac{q}{x+q} \ln(q)$  with  $q > 0$  and domain  $A = (0, +\infty)$ . Notice that  $f_q(q) = 0$ . We will prove that  $f_q$  is strictly increasing for all  $x \in (0, q]$ , which for  $0 < p < q$  gives the inequality  $f_q(p) < f_q(q)$ . Then, it is easy to see that the latter is equivalent to (#). Therefore let us compute the first derivative of  $f$ :

$$\begin{aligned} f'_q(x) &= \frac{1}{2} \cdot \frac{2}{x+q} - \frac{(x+q) - x}{(x+q)^2} \ln(x) - \frac{x}{x+q} \cdot \frac{1}{x} + \frac{q}{(x+q)^2} \ln(q) = \frac{1}{x+q} - \frac{q \ln(x)}{(x+q)^2} - \frac{1}{x+q} + \frac{q \ln(q)}{(x+q)^2} \\ &= \frac{q(\ln(q) - \ln(x))}{(x+q)^2} = \frac{q \ln\left(\frac{q}{x}\right)}{(x+q)^2}. \end{aligned}$$

The critical points of  $f_q$  are the solutions of  $f'_q(x) = 0$ , and we see that

$$f'_q(x) = 0 \iff \ln\left(\frac{q}{x}\right) = 0 \iff \frac{q}{x} = 1 \iff q = x.$$

Now, for  $x > q$  we have  $\ln\left(\frac{q}{x}\right) < 0$ , hence  $f'_q(x) < 0$  and  $f_q$  is strictly decreasing in  $[q, +\infty)$  for all  $q > 0$ . On the other hand for  $0 < x < q$  we get  $f'_q(x) > 0$ , therefore  $f_q$  is strictly increasing on  $(0, q]$  for all  $q > 0$ . Thus, in fact, at  $x_0 = q$  our family  $f_q$  attains its unique maximum which equals to  $f_q(q) = 0$  (hence,  $f_q$  is non-positive all along its domain, for all  $q > 0$ ). Let us finally sketch the graph of  $f_q$  for  $q = 1, 2, \dots, 5$ , so that we have an illustrated confirmation of the monotonicity of  $f_q$  for these values of the parameter  $q$ .



To obtain this figure, we have used Sage and a bit of programming (relying on the command `for`), which goes as follows:

```
var("q"); f(x, q)=ln((x+q)/2)-(x/(x+q))*ln(x)-(q/(x+q))*ln(q)
a=plot([f(x, q) for q in [1,2,..,5]], (x, 0, 10), ymax=0.05, ymin=-0.4)
a+=point([(q, f(q, q)) for q in [1,2,..,5]], size=20, color="black")
a+=text(r"$q=1$", (1, 0+0.02), fontsize=12, color="black")
a+=text(r"$q=2$", (2, 0+0.02), fontsize=12, color="black")
a+=text(r"$q=3$", (3, 0+0.02), fontsize=12, color="black")
a+=text(r"$q=4$", (4, 0+0.02), fontsize=12, color="black")
a+=text(r"$q=5$", (5, 0+0.02), fontsize=12, color="black"); show(a)
```

□

When applying the l'Hopital's rule, caution is always necessary. For instance, the l'Hopital's rule can yield a non-existent limit even when the original limit exists. Let's illustrate this remarkable scenario with an example.

**5.E.108.** Consider the function  $f(x) = \frac{x + \sin x}{x}$ .

(a) Which is the type of indeterminate form corresponding to the limit  $\lim_{x \rightarrow +\infty} f(x)$ ?

(b) Show that in this case the l'Hopital's rule leads to a non-existing limit, although  $\lim_{x \rightarrow +\infty} f(x) = 1$ . ○

**5.E.109.** Using the L'Hospital's rule provide an alternative proof of the relation  $\lim_{n \rightarrow +\infty} \sqrt[n]{n} = 1$  (see 5.B.8). Moreover, apply the same method to prove that  $\lim_{n \rightarrow +\infty} \sqrt[n]{\ln(n)} = 1$  (see 5.E.18), and  $\lim_{n \rightarrow +\infty} (n \sin(\frac{1}{n})) = 1$ .

**Solution.** Recall that  $\sqrt[n]{n} = n^{1/n}$ . Thus, as  $n \rightarrow +\infty$  this gives the indeterminate form  $(+\infty)^0$ . Consider the function  $f(x) = x^{1/x}$  with  $x > 0$ . Taking the logarithm in both sides we get  $\ln(f(x)) = \ln(x^{1/x}) = \frac{1}{x} \ln(x)$ . Thus  $f(x) = e^{\frac{\ln(x)}{x}}$ .

Notice that by the L'Hospital's rule we get  $\lim_{x \rightarrow +\infty} \frac{\ln(x)}{x} \stackrel{\frac{+\infty}{+\infty}}{=} \lim_{x \rightarrow +\infty} \frac{1}{x} = 0$ . Therefore

$$\lim_{x \rightarrow +\infty} f(x) = \lim_{x \rightarrow +\infty} e^{\frac{\ln(x)}{x}} = e^{\lim_{x \rightarrow +\infty} \frac{\ln(x)}{x}} = e^0 = 1,$$

and applying this to our sequence we get the result:

$$\lim_{n \rightarrow +\infty} \sqrt[n]{n} = \lim_{n \rightarrow +\infty} n^{\frac{1}{n}} = \lim_{n \rightarrow +\infty} e^{\frac{\ln(n)}{n}} = 1.$$

Let us use the L'Hospital's rule to prove also that  $\lim_{n \rightarrow +\infty} \sqrt[n]{\ln(n)} = 1$  and leave the third limit for practice. Since we have  $\sqrt[n]{\ln(n)} = (\ln(n))^{\frac{1}{n}}$ , we see that the limit  $\lim_{n \rightarrow +\infty} \sqrt[n]{\ln(n)} = 1$  has the indeterminate form  $(+\infty)^0$ . So, let us consider the function  $g(x) = (\ln(x))^{\frac{1}{x}}$ . Then, we can write  $g(x) = e^{\frac{1}{x} \cdot \ln(\ln(x))}$ . Hence, to compute  $\lim_{x \rightarrow +\infty} g(x)$  it is sufficient to compute  $\lim_{x \rightarrow +\infty} \frac{\ln(\ln(x))}{x}$ . This has the indeterminate form  $\frac{+\infty}{+\infty}$ , so by applying the L'Hospital's rule we get

$$\lim_{x \rightarrow +\infty} \frac{\ln(\ln(x))}{x} = \lim_{x \rightarrow +\infty} \frac{(\ln(\ln(x)))'}{x'} = \lim_{x \rightarrow +\infty} \frac{\frac{(\ln(x))'}{\ln(x)}}{1} = \lim_{x \rightarrow +\infty} \frac{1}{x \ln(x)} = \frac{1}{+\infty} = 0.$$

Therefore,  $\lim_{x \rightarrow +\infty} g(x) = e^0 = 1$  and this implies that  $\lim_{n \rightarrow +\infty} \sqrt[n]{\ln(n)} = 1$ . □

**5.E.110.** Use the l'Hospital's rule to evaluate the limit  $\lim_{x \rightarrow 0} (\cot(x) - \frac{1}{x})$ .

**Solution.** Since  $\cot(x) = \frac{\cos(x)}{\sin(x)}$ , by applying repeatedly the l'Hospital's rule one has

$$\begin{aligned} \lim_{x \rightarrow 0} \left( \cot(x) - \frac{1}{x} \right) &= \lim_{x \rightarrow 0} \frac{x \cos(x) - \sin(x)}{x \sin(x)} \stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0} \frac{(x \cos(x) - \sin(x))'}{(x \sin(x))'} = \lim_{x \rightarrow 0} \frac{\cos(x) - x \sin(x) - \cos(x)}{\sin(x) + x \cos(x)} \\ &= \lim_{x \rightarrow 0} \frac{-x \sin(x)}{\sin(x) + x \cos(x)} \stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0} \frac{(-x \sin(x))'}{(\sin(x) + x \cos(x))'} = \lim_{x \rightarrow 0} \frac{-\sin(x) - x \cos(x)}{\cos(x) + \cos(x) - x \sin(x)} \\ &= \frac{0 - 0}{1 + 1 - 0} = 0. \end{aligned}$$

□

**5.E.111.** Using the l'Hospital's rule, when adequate, determine the limits

$$a = \lim_{x \rightarrow 1^-} (1 - x) \tan\left(\frac{\pi x}{2}\right), \quad b = \lim_{x \rightarrow \frac{\pi}{2}^-} \left(\frac{\pi}{2} - x \tan(x)\right), \quad c = \lim_{x \rightarrow 1} \left(\frac{1}{2 \ln(x)} - \frac{1}{x^2 - 1}\right), \quad d = \lim_{x \rightarrow +\infty} \left(\cos \frac{2}{x}\right)^{x^2}.$$

○

Next we present a result which analogous to l'Hopital's rule, which can be used in establishing the monotonicity of a ratio  $f/g$  of two functions  $f, g : [a, b] \rightarrow \mathbb{R}$ , which are both differentiable on  $(a, b)$ , with  $g'(x) \neq 0$  for all  $x \in (a, b)$ . In the literature, it is commonly referred to as the "l'Hopital's Monotone Rule" or for short by "LMR", and, despite its numerous applications in various contexts, it often seems to be overlooked. Notice the rule holds true even in cases where  $a$  or  $b$  is infinite, and in particular when  $f, g$  are defined only on the open interval  $(a, b)$  (provided that  $g'$  has constant sign).

**5.E.112.** Suppose that  $a, b$  are such that  $-\infty \leq a < b \leq \infty$  and let  $f, g : (a, b) \rightarrow \mathbb{R}$  be continuous functions which are differentiable on  $(a, b)$ . Suppose also that  $\lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a^+} g(x) = 0$  (or  $\lim_{x \rightarrow b^-} f(x) = \lim_{x \rightarrow b^-} g(x) = 0$ ), and that the derivative  $g'$  is nonzero (and does not change sign on  $(a, b)$ ). Prove that if  $f'/g'$  is (strictly) increasing or (strictly) decreasing on  $(a, b)$ , then so is  $f/g$ .

**Solution.** We will analyze the case with  $\lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a^+} g(x) = 0$ , and similarly one can treat the case  $f(b^-) = g(b^-) = 0$ . Assume also that  $f'/g'$  is strictly increasing, and the proof for the rest cases is analogous. We will prove that

$$\left(\frac{f}{g}\right)' > 0, \quad \text{for all } x \in (a, b),$$

and hence the ratio  $f/g$  is strictly increasing. Fix some arbitrary  $x \in (a, b)$  and consider the function  $h_x : (a, b) \rightarrow \mathbb{R}$ , defined by

$$h_x(y) = f'(x)g(y) - f(y)g'(x), \quad y \in (a, b).$$

Obviously,  $h_x$  is differentiable on  $(a, b)$  (and hence also continuous), with

$$h'_x(y) = f'(x)g'(y) - f'(y)g'(x) = g'(x)g'(y) \left(\frac{f'(x)}{g'(x)} - \frac{f'(y)}{g'(y)}\right). \quad (*)$$

By the hypothesis,  $g'(x)$  is non-zero and it has constant sign, that is, only one of the following relations is possible:  $g'(x) > 0$  or  $g'(x) < 0$  for all  $x \in (a, b)$ . Moreover  $f'/g'$  is strictly increasing, so for  $x > y$  the inequality  $\frac{f'(x)}{g'(x)} > \frac{f'(y)}{g'(y)}$  makes sense. Therefore, by (\*) we deduce that  $h'_x(y) > 0$  for all  $y \in (a, x)$ , which means that the function  $h_x$  is strictly increasing on  $(a, x)$ . In fact, by continuity,  $h_x$  should be strictly increasing on  $(a, x]$ . Moreover, by our hypothesis we get

$$\lim_{x \rightarrow a^+} h_x(x) = \lim_{x \rightarrow a^+} (f'(x)g(x) - f(x)g'(x)) = 0,$$

and hence, all together, this implies that  $h_x(x) > 0$ . Thus

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)} = \frac{h_x(x)}{g^2(x)} > 0$$

and since  $x \in (a, b)$  is arbitrary, our claim follows.  $\square$

**5.E.113.** (a) Based on the l'Hopital monotone rule study the monotonicity of  $k = f/g$  on the specified domain  $A$ , when  $f, g, A$  are given as follows:

(1)  $f(x) = e^{2x} - 1, g(x) = \ln(x + 1), A = (-1, 0) \cup (0, +\infty)$ ;

(2)  $f(x) = e^x - 1, g(x) = 2x + e^x - 1, A = (-\infty, 0) \cup (0, \infty)$ .

Then, confirm your conclusions by specifying the sign of the first derivative  $\kappa'$  of  $\kappa = f/g$ , and rely on Sage to verify all your computations.

(b) For both cases confirm that  $\lim_{x \rightarrow 0} \kappa(x) = \lim_{x \rightarrow 0} K(x)$ , where  $K(x) = f'(x)/g'(x)$ .

**Solution.** (a) Let us begin with (1). We have

$$\kappa(x) = \frac{f(x)}{g(x)} = \frac{e^{2x} - 1}{\ln(x + 1)}, \quad \text{and} \quad K(x) = \frac{f'(x)}{g'(x)} = 2(x + 1)e^{2x},$$

where  $\kappa$  is defined on  $A = (-1, 0) \cup (0, +\infty)$  (and we assume the same for both  $f, g$ ). Moreover, the given functions satisfy  $\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^+} g(x) = 0$ , hence one can apply the LMR (5.E.112). A direct computation shows that

$$K'(x) = 2e^{2x}(3 + 2x) > 0, \quad \text{for all } x \in A.$$

Because  $K = f'/g'$  is differentiable, this implies that  $K$  strictly increases on  $A$  and according to LMR the same monotonicity pattern we should get for the function  $\kappa = f/g$ . To confirm this result, and because  $\kappa$  is differentiable on  $A$ , we will use its first derivative:

$$\frac{d\kappa(x)}{dx} = \kappa'(x) = \left(\frac{f(x)}{g(x)}\right)' = \left(\frac{e^{2x} - 1}{\ln(x + 1)}\right)' = \frac{2e^{2x} \cdot \ln(x + 1) - (e^{2x} - 1) \cdot \frac{1}{x+1}}{(\ln(x + 1))^2} = \frac{2(x + 1)e^{2x} \ln(x + 1) - e^{2x} + 1}{(x + 1)(\ln(x + 1))^2}.$$

To determine the sign of  $\kappa'(x)$  it is sufficient to determine the sign of the numerator  $h(x) := 2(x + 1)e^{2x} \ln(x + 1) - e^{2x} + 1$ . A direct computation shows that  $h'(x) = 2e^{2x} \ln(x + 1)(2x + 3)$ , so for all  $x \in A$  we have  $h'(x) > 0$ . Therefore  $h$  is increasing on  $A$ . Moreover  $h(0) = 0$ , and thus it should be  $h(x) > 0$  for all  $x \in A$ , which implies that  $\kappa'(x) > 0$ . As a side remark observe that it is easier to deduce that the ratio  $K = f'/g'$  is increasing, than determining the sign of  $k' = (f/g)'$ .

Now, let's perform all computations in Sage, including necessary explanations within the code:

```
f(x)=e^(2*x)-1; g(x)=ln(x+1); k(x)=f(x)/g(x) #declare the ratio k=f/g
A=RealSet((-1, 0), x>0); show(A) #declare the domain A of k
K(x)=diff(f, x)/diff(g, x); show(K(x)) #declare and show the ratio K=f'/g'
Kprime(x)=diff(K, x).factor(); show(Kprime(x)) #declare and show the derivative K'
```

```

print(bool(Kprime(x)>0 for x in A)) #confirm that K'>0 for all the elements of A
kprime(x)=diff(k, x).factor(); show(kprime(x)) #declare and show the derivative k'
print(bool(kprime(x)>0 for x in A)) #confirm that k'>0 for all the elements of A
h(x)=2*(x+1)*e^(2*x)*ln(x+1)-e^(2*x)+1; show(diff(h, x).factor()) #declare the numerator h
print(bool(diff(h, x)>0 for x in A)) #confirm that the derivative of h is positive in A
print(bool(h(x)>0 for x in A)) #confirm that h is positive in A

```

(2) In this case we have

$$\kappa(x) = \frac{f(x)}{g(x)} = \frac{e^x - 1}{2x + e^x - 1}, \quad \text{and} \quad K(x) = \frac{f'(x)}{g'(x)} = \frac{e^x}{2 + e^x},$$

where  $\kappa$  is defined on  $A = (-\infty, 0) \cup (0, +\infty)$  (and we assume the same for both  $f, g$ ). We also have  $\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^+} g(x) = 0$  and hence we can apply the LMR. In particular, the ratio  $K(x) = \frac{e^x}{2+e^x}$  is strictly increasing on  $A$  (e.g., its first derivative satisfies  $K'(x) = \frac{2e^x}{(2+e^x)^2} > 0$  for all  $x \in A$ ), and hence the same monotonicity pattern must have the ratio  $\kappa = f/g$ . It is an easy exercise to see that the first derivative of  $\kappa$  satisfies

$$\kappa'(x) = \frac{2(xe^x - 1 - e^x + 1)}{(2x + e^x - 1)^2} > 0, \quad \forall x \in A.$$

On the other hand, a confirmation via Sage occurs in a very similar way as in (a), i.e.,

```

f(x)=e^x-1; g(x)=2*x+e^x-1; k(x)=f(x)/g(x)
A=RealSet(x>0, x<0); show(A) #declare the domain A
K(x)=diff(f, x)/diff(g, x); show(K(x)) #declare and show the ratio K=f'/g'
Kprime(x)=diff(K, x).factor(); show(Kprime(x)) #declare and show the derivative K'
print(bool(Kprime(x)>0 for x in A)) #confirm that K'>0 for all the elements of A
kprime(x)=diff(k, x).factor(); show(kprime(x)) #declare and show the derivative k'
print(bool(kprime(x)>0 for x in A)) #confirm that k'>0 for all the elements of A

```

(b) The objective of this question is to illustrate that the LMR operates similarly to the l'Hospital's rule for limits. According to the latter rule we get

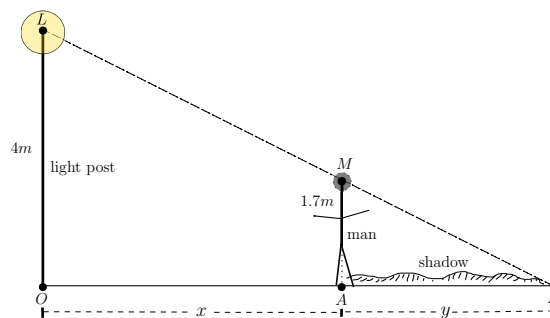
$$\lim_{x \rightarrow 0} \kappa(x) = \lim_{x \rightarrow 0} \frac{f(x)}{g(x)} = \lim_{x \rightarrow 0} \frac{e^{2x} - 1}{\ln(x + 1)} \stackrel{\frac{0}{0}}{=} \lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)} = \lim_{x \rightarrow 0} K(x) = \lim_{x \rightarrow 0} 2(x + 1)e^{2x} = 2.$$

Verify yourself that in a similar way in (2) we get  $\lim_{x \rightarrow 0} \kappa(x) = \lim_{x \rightarrow 0} K(x) = 1/3$ . □

C2) MATERIAL ON DERIVATIVES - APPLICATIONS

**5.E.114. Thales theorem and the shadow problem.** (a) A man who is 1.7 meters tall walks away from a 4-meter-tall lamppost at a speed of 1.4 m/sec (see also the figure given below). Find the rate at which his shadow is increasing in length. (b) Suppose now that as a man walks away from a lamppost with height 380 cm, the tip of his shadow moves twice as fast as he does. What is man's height?

**Solution.** (a) Let us denote by  $x$  the distance of the man from the base of the lamppost at time  $t$ , and by  $y$  the length of his shadow, both measured as meters. By assumption the man is moving at a rate of 1.4m/sec, that  $x = 1.4t$ . Consider the triangles  $OLB$  and  $AMB$ .



These triangles are similar, so by Thales theorem we have

$$\frac{x + y}{4} = \frac{y}{1.7} \iff \frac{1.4t + y}{4} = \frac{y}{1.7} \iff y = 1.03478 \cdot t = \frac{119}{115} \cdot t.$$

One may like to solve this equation via Sage and take the final equality presented here, which can be done by the syntax

```
var("t"); solve(4*x==1.7*(1.4*t+x), x)
```

It follows that the length of man's shadow increases with rate  $\frac{dy}{dt} = \frac{119}{115}$  m/sec.

(b) Let us denote by  $\alpha$  the whole distance from lamppost's base to the tip of man's shadow (hence in terms of the previous figure we have  $\alpha = x + y$ ), and by  $h$  the man's height (which we now enumerate in centimeters and is the unknown of the problem). Notice that both  $\alpha, x$  are functions of the time  $t$ , but  $h$  is constant. In this case Thales theorem tells us that

$$\frac{\alpha(t)}{380} = \frac{\alpha(t) - x(t)}{h} \iff 380 \cdot (\alpha(t) - x(t)) = h \cdot \alpha(t)$$

which we need to differentiate with respect to  $t$ , based on the statement of the assumption, that is  $\frac{d\alpha}{dt} = 2\frac{dx}{dt}$ . This gives

$$380\left(\frac{d\alpha}{dt} - \frac{dx}{dt}\right) = 2 \cdot h \frac{dx}{dt} \iff 380\left(2\frac{dx}{dt} - \frac{dx}{dt}\right) = 2 \cdot h \frac{dx}{dt},$$

which is equivalent to  $h = \frac{380}{2} = 190$  cm. □

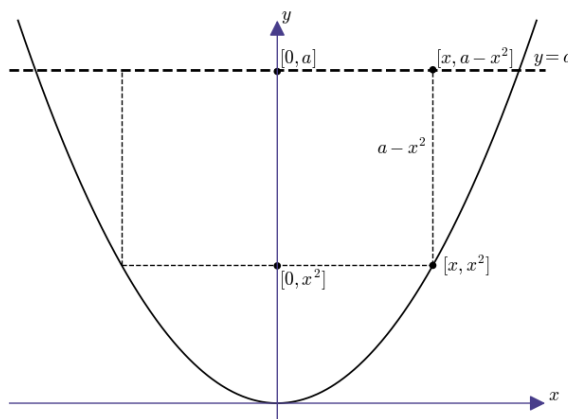
**5.E.115. Inscribed rectangle with maximal area.** (1) Find the rectangle with the largest area that can be inscribed inside the graph of the parabola  $y = x^2$  and below the line  $y = a$ , where  $a > 0$  is some fixed constant (see the figure below).

(2) Which is the maximum area for the values 1, 2 and 3 of  $a$ ?

**Solution.** (1) Let  $[x, x^2]$  be the point representing the lower right corner of the rectangle. As we indicate in the figure, the upper right corner should be the point  $[x, a]$ , in particular the height of the rectangle is the function  $h_a(x) = a - x^2$  with  $x \in [0, \sqrt{a}]$ . Notice  $h_a(\sqrt{a}) = 0$ , and  $h_a(0) = a$  which are two cases that we may exclude for obvious reasons (these correspond to zero height and zero width ( $x = 0$ ) respectively, hence they do not determine a rectangle). Thus, the function representing the area of the rectangle is

$$E_a(x) = 2 \cdot x \cdot h_a(x) = 2x(a - x^2) = 2xa - 2x^3,$$

which we want to maximize on the interval  $(0, \sqrt{a})$ .



A differentiation of the defining equation of  $E_a$  with respect to  $x$  gives  $E'_a(x) = 2a - 6x^2$ . The equation  $E'_a(x) = 0$  has a unique solution in  $(0, \sqrt{a})$  given by  $x_a := \sqrt{\frac{a}{3}}$ , with

$$E_a(x_a) = \frac{4\sqrt{3}}{9}a^{\frac{3}{2}} = \frac{4\sqrt{3}a^3}{9}, \quad a > 0. \quad (*)$$

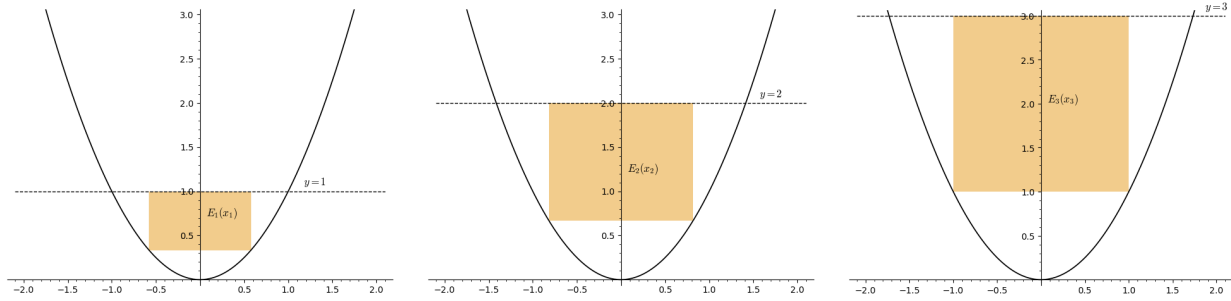
There are many ways to prove that  $x_a$  maximizes the area. For instance the second derivative of the area function  $E_a$  is given by  $E''_a(x) = (E'_a(x))' = -12x$  and hence obviously  $E''_a(x_a) < 0$  (since  $x_a > 0$  for all  $a > 0$ ). You can easily confirm these conclusions in Sage by applying the cell

```
var("a"); assume(a>0)
h(a, x)=a-x^2; E(a, x)=2*x*h(a, x)
show(diff(E(a, x), x)); show(solve(diff(E(a, x), x)==0, x))
bool(diff(E(a, x), x, 2)(x=sqrt(a/3))<0)
```

(2) It is sufficient to replace in (\*) the required values. This gives  $E_1(x_1) = \frac{4\sqrt{3}}{9}$ ,  $E_2(x_2) = \frac{4\sqrt{24}}{9} = \frac{8\sqrt{6}}{9}$  and  $E_3(x_3) = \frac{4\sqrt{81}}{9} = 4$ . In Sage to confirm these values add in the previous block the cell

```
show(E(1, sqrt(1/3)).factor()); show(E(2, sqrt(2/3)).factor()); show(E(3, sqrt(3/3)).factor())
```

Let us finalize this task by illustrating these three solutions:



□

**5.E.116.** Inscribe a rectangle with the greatest perimeter possible into a semidisc with radius  $r$ . Determine the rectangle's perimeter.

○

**5.E.117.** Among the rectangles with perimeter  $4c$ , find the one having the greatest area (if such one exists) and determine the lengths of its sides.

○

**5.E.118.** Find the height  $h$  and the radius  $r$  of the largest (greatest volume) cone which fits into a ball of radius  $R$ .

○

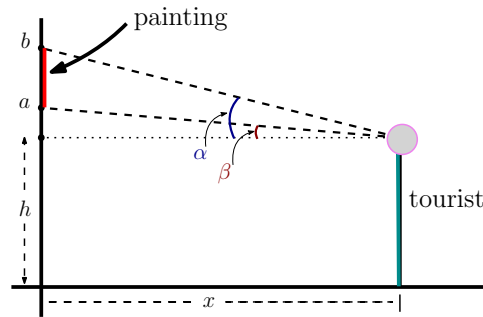
**5.E.119.** From all triangles with given perimeter  $p$ , select the one with the greatest area.

○

**5.E.120.** A parabola is given by the equation  $2x^2 - 2y = 9$ . Find the points of the parabola which are closest to the origin.

○

**5.E.121. Regiomontanus' angle maximization problem (1471).** In the museum, there is a painting on the wall. Its lower edge is  $a$  meters above the ground and its upper edge is  $b$  meters, that is, its height equals to  $b - a$ . A tourist is looking at the painting, her eyes being at height  $h < a$  meters above ground (see the figure below). How far from the wall should the tourist stand if she wants to maximize her angle of view at the painting? This is the so-called "Regiomontanus' angle maximization problem" from 14th century.<sup>19</sup>



**Solution.** First observe that because the tourist is below the lowest level of the painting, her viewing angle very close to the wall of very far from the wall becomes small, so the largest viewing angle is somewhere in between. We will use the picture above (at the left) to illustrate the situation. Let us denote by  $x$  the distance (in meters) of the tourist from the wall and the angle of her view at the painting by  $\varphi$ . For the angles  $\alpha, \beta \in (0, \pi/2)$  we have

$$\tan(\alpha) = \frac{b-h}{x}, \quad \tan(\beta) = \frac{a-h}{x},$$

with  $h < a < b$ . Our task is to maximize  $\varphi = \alpha - \beta$ . Let us add that for  $h > b$ , one can proceed analogously and for  $h \in [a, b]$ , the angle  $\varphi$  increases as  $x$  decreases ( $\varphi = \pi$  for  $x = 0$  and  $h \in (a, b)$ ). From the condition  $h < a$  it follows that the angle  $\varphi$  is acute, that is,  $\varphi \in (0, \pi/2)$ . Since the function  $y = \tan(x)$  is increasing on the interval  $(0, \pi/2)$ , we can turn our attention on maximizing the value  $\tan(\varphi)$ . We have that

<sup>19</sup> This optimization problem was first posed and solved in 1471 by the German astronomer Johannes Müller von Königsberg (1436–1476), known as Regiomontanus. Müller solved the problem based on elementary geometry, but the solution presented here relies on differentiation.

$$\tan(\varphi) = \tan(\alpha - \beta) = \frac{\tan(\alpha) - \tan(\beta)}{1 + \tan(\alpha)\tan(\beta)} = \frac{\frac{b-h}{x} - \frac{a-h}{x}}{1 + \frac{b-h}{x} \cdot \frac{a-h}{x}} = \frac{x(b-a)}{x^2 + (b-h)(a-h)}.$$

So it suffices to find the global maximum of the function

$$f(x) = \frac{x(b-a)}{x^2 + (b-h)(a-h)}, \quad x \in [0, +\infty),$$

whose first derivative is given by

$$f'(x) = \frac{(b-a)[(b-h)(a-h) - x^2]}{[x^2 + (b-h)(a-h)]^2}, \quad (*)$$

for all  $x \in (0, +\infty)$ . This expression can be quickly verified via the following cell in Sage:

```
var("a, b, h"); f(x)=x*(b-a)/(x^2+(b-h)*(a-h))
df1=diff(f(x), x).factor(); show(df1)
```

From (\*) one can see that

$$f'(x) > 0, \quad \text{for } x \in (0, \sqrt{(b-h)(a-h)}), \quad \text{and } f'(x) < 0, \quad \text{for } x \in (\sqrt{(b-h)(a-h)}, +\infty),$$

and hence  $f$  attains its global maximum at the point  $x_0 = \sqrt{(b-h)(a-h)}$ . This is the desired point and the corresponding angle is given by

$$\varphi_0 = \arctan\left(\frac{x_0(b-a)}{x_0^2 + (b-h)(a-h)}\right) = \arctan\left(\frac{b-a}{2\sqrt{(b-h)(a-h)}}\right).$$

Let us describe two examples:

- Suppose an ant looks at the painting, so in this case we have  $h = 0$ ,  $x_0 = \sqrt{ab}$  and  $\varphi_0 = \arctan\left(\frac{b-a}{2\sqrt{ab}}\right)$ . If the painting is 1 meter high and its lower edge is 2 meters above ground, i.e.,  $a = 2$  and  $b = 3$ , then the ant will see the painting at the largest angle  $\varphi_0 \doteq 0.201\text{rad} \approx 11.5^\circ$  at the distance  $x_0 = \sqrt{6} \approx 2.45$  meters from the wall.

- As for another example, look in the left picture above, where the painting is viewed by a man whose eyes are at the height of 1.8 meters, together with his son whose eyes are 1 meter above ground. In this case the father should stand at  $x_0 \doteq 0.49$  meters from the wall, and his son at  $x_0 \doteq 1.41$  meters from the wall. Thus the father has viewing angle  $\varphi_0 \doteq 0.795\text{rad} \approx 45.6^\circ$  whereas his son has  $\varphi_0 \doteq 0.339\text{rad} \approx 19.5^\circ$ , and the quotient  $\frac{0.795}{0.339} \approx 2.3$  proves what a strongly better view the father has. □

**5.E.122.** Determine the point  $x_0$  of the previous problem 5.E.121 by an alternative method. ○

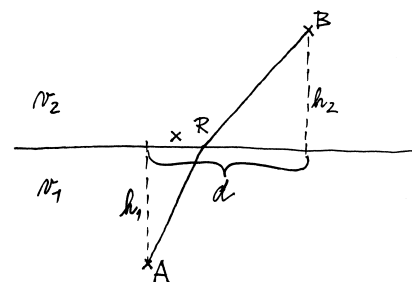
**5.E.123. Snell's law.** This task is about the light rays which bend when traveling from one (optical) medium to another, such as air to water. This causes items placed in or behind water to look shifted and there is a rule which can be applied to evaluate the bending of a light ray, called the “*Snell's law*”.<sup>20</sup> Having as reference the figure given below, determine the refracted light ray between the point  $A$  in a homogeneous medium with speed of light  $v_1$  and the point  $B$  in a homogeneous medium with speed of light  $v_2$ .



**Solution.**

We can assume that distances are given in meters, the speeds  $v_1, v_2$  in meters per second, and the time in seconds. But for convenience we will omit the units of measurement in our notation. The ray is determined by “*Fermat's principle of least time*”, which states that “from all the paths between the points  $A$  and  $B$ , the light will go along the one which can be traversed in the least time”. In homogeneous mediums the ray will be a straight line and in our case we will consider its segment. So it suffices to determine the point  $R$  (given by the value  $x$ ) where the ray refracts. The distance between the points  $A$  and  $R$  is  $\sqrt{h_1^2 + x^2}$ , while the distance between  $R$  and  $B$  is  $\sqrt{h_2^2 + (d-x)^2}$ . We deduce that the total time which needs the energy to be transmitted between the points  $A$  and  $B$  is represented by the function

$$T(x) = \frac{\sqrt{h_1^2 + x^2}}{v_1} + \frac{\sqrt{h_2^2 + (d-x)^2}}{v_2},$$



<sup>20</sup>An optical medium is material through which light (and other electromagnetic waves) propagate.

with  $x \in [0, d]$ . We want to locate the point  $x \in [0, d]$  at which the value  $T(x)$  is minimal. The derivative of  $T$  is the continuous function  $T'(x) = \frac{x}{v_1 \sqrt{h_1^2 + x^2}} - \frac{d-x}{v_2 \sqrt{h_2^2 + (d-x)^2}}$  for all  $x \in [0, d]$ , and thus the critical point equation

$T'(x) = 0$  is equivalent to  $\frac{x}{v_1 \sqrt{h_1^2 + x^2}} = \frac{d-x}{v_2 \sqrt{h_2^2 + (d-x)^2}}$ , from where we deduce that

$$\frac{f_1(x)}{f_2(x)} = \frac{v_1}{v_2}, \quad \text{with} \quad f_1(x) = \frac{x}{\sqrt{h_1^2 + x^2}}, \quad f_2(x) = \frac{d-x}{\sqrt{h_2^2 + (d-x)^2}}.$$

This expression is useful for us because (see the picture)

$$\sin(\varphi_1) = \frac{x}{\sqrt{h_1^2 + x^2}}, \quad \sin(\varphi_2) = \frac{d-x}{\sqrt{h_2^2 + (d-x)^2}}.$$

Thus there is at most one stationary point, determined by the relation

$$\frac{\sin(\varphi_1)}{\sin(\varphi_2)} = \frac{v_1}{v_2}. \quad (\#)$$

This is the so called “Snell’s law” in physics. As  $\varphi_1 \in [0, \pi/2]$  increases the angle  $\varphi_2 \in [0, \pi/2]$  decreases. On the other hand, on the interval  $[0, \pi/2]$  the sine function is non-negative and increasing, so the quotient  $(\sin \varphi_1)/(\sin \varphi_2)$  is increasing with respect to  $x$ . Since  $T'(0) < 0$  and  $T'(d) > 0$ , there is exactly one stationary point  $x_0$ . From the inequalities  $T'(x) < 0$  for  $x \in [0, x_0)$  and  $T'(x) > 0$  for  $x \in (x_0, d]$ , it follows that at the stationary point  $x_0$  there is the global minimum.

Let us summarize as follows: The ray is given by the point  $R$  of refraction (i.e., the value  $x_0$ ), and the point  $R$  is determined by the Snell’s law (#). The ration  $v_1/v_2$  is constant for the given homogeneous mediums, and determines an important quantity which describes the interface of optical spaces. It is called “refractive index” and denoted by  $\mathbf{n}$ . Usually, the first optical medium (space) is vacuum, that is,  $v_1 = c$ , where  $c =$  speed of light and  $v_2 = v$ , such that  $\mathbf{n} = c/v$ . For vacuum, we get  $\mathbf{n} = 1$ , of course. This value is also used for the air since its refractive index at standard conditions is  $\mathbf{n} \doteq 1.000272$ . Other mediums have  $\mathbf{n} > 1$  ( $\mathbf{n} = 1.31$  for ice,  $\mathbf{n} = 1.33$  for water,  $\mathbf{n} = 1.5$  for glass). However, the refractive index also depends on the wave length of the electromagnetic radiation in question (for example, for water and light, it ranges from  $\mathbf{n} \doteq 1.331$  to  $\mathbf{n} \doteq 1.344$ ), where the index ordinarily decreases as the wave length increases. The speed of light in an optical space having  $\mathbf{n} > 1$  depends on its frequency, and hence we talk about the “dispersion of light”. The dispersion causes rays of different colors to refract at different angles. The violet ray refracts the most and the red ray refracts the least. This is also the origin of a rainbow.  $\square$

**5.E.124.** You are in a boat on a lake at distance  $d$  km from the shore. You want to get to a given place on the shore whose straight-line distance is  $\sqrt{d^2 + \ell^2}$  from you, see the figure below. What path will you take if you want to be there as soon as possible, supposing you can row at  $v_1$  kph and run along the shore at  $v_2$  kph? How long will the journey take?

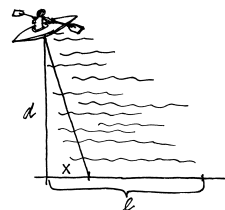


**Solution.** The optimal strategy is given by first rowing in a straight line to the shore at some point  $[0, x]$  for  $x \in [0, \ell]$ , and then running along the shore to the target point  $[0, \ell]$ , see the diagram. So the trajectory consists of two line segments, or only one segment, in the case when  $x = \ell$ . The voyage to the point  $[0, x]$  on the shore takes  $\frac{\sqrt{d^2+x^2}}{v_1}$  hours, and the final run takes  $\frac{\ell-x}{v_2}$  hours. Thus, the total time is the function

$$t(x) = \frac{\sqrt{d^2 + x^2}}{v_1} + \frac{\ell - x}{v_2}, \quad x \in [0, \ell].$$

It can be assumed that  $v_1 < v_2$ , for if  $v_1 \geq v_2$ , the optimal strategy is to row straight to the target point, which corresponds to  $x = \ell$ . Now, the first and the second derivative of  $t(x)$  are given by  $t'(x) = \frac{x}{v_1 \sqrt{d^2+x^2}} - \frac{1}{v_2}$  and  $t''(x) = \frac{d^2}{v_1 \sqrt{(d^2+x^2)^3}}$ , respectively, with  $x \in (0, \ell)$ . You may confirm these relations in Sage as follows:

```
var("d, l, v1, v2"); assume(v1<v2)
t(x)=(sqrt(d^2+x^2))/(v1) + (l-x)/(v2)
d1t(x)=diff(t, x); show(d1t(x))
d2t(x)=diff(t, x, 2); show(d2t(x).factor())
```



In this block for the second derivative we used the command `diff(t, x, 2)`, where recall that `diff(f, x, k)` computes the  $k$ th derivative of some function  $f$  (see also Chapter 6). As an alternative one could simply type `d2t(x) = diff(d1t, x)`, since in our code `d1t` represents the first derivative of  $t(x)$ .



Let us now focus on the stationary point equation  $t'(x) = 0$ , or equivalently  $\frac{x}{\sqrt{d^2+x^2}} = \frac{v_1}{v_2}$ , whose solution is given by

$$x_0 = \frac{v_1 d}{\sqrt{v_2^2 - v_1^2}}.$$

You can confirm this expression via Sage by adding in the previous block the command

```
show(solve(d1t(x)==0, x))
```

We see that if  $x_0 < \ell$ , then  $t(x)$  has a global minimum at  $x_0$  on the interval  $[0, \ell]$ . This is because  $t''(x) > 0$  for all  $x$  in that interval. If  $x_0 \geq \ell$ , then  $t'(x) \leq 0$  for all  $x \in [0, \ell]$  and so the global minimum of  $t(x)$  occurs at  $\ell$ . In the former case, the fastest journey in hours takes

$$t(x_0) = \frac{\sqrt{d^2 + x_0^2}}{v_1} + \frac{\ell - x_0}{v_2} = \frac{d \sqrt{v_2^2 - v_1^2}}{v_1 v_2} + \frac{\ell}{v_2}.$$

In the latter case, the fastest journey in hours takes  $t(\ell) = \frac{\sqrt{d^2 + \ell^2}}{v_1}$ .

Notice one could follow an even simpler approach for doing the calculations, using the variable  $\theta$ , instead of the variable  $x$ , where  $x = d \tan(\theta)$ . The fastest journey occurs when  $\sin(\theta) = v_1/v_2$ , which is the limiting case of Snell's law.  $\square$

### 5.E.125. L'Hospital's pulley.

A rope of length  $r$  is tied at one of its ends to the ceiling at a point  $A$ . A small pulley is attached to its other end. A point  $B$  is also on the ceiling at distance  $d$ ,  $d > r$ , from  $A$ . Another rope of length  $\ell > \sqrt{d^2 + r^2}$ , is tied to  $B$  at one end, passes over the pulley, and has a weight is attached to its other end, see the figure presented below. Omit the mass and the size of the ropes and the pulley. In what position does the weight stabilize so that the system is in a stationary position?



**Solution.** The system is in stationary position when its potential energy is minimized. This is when the distance between the weight and the ceiling is maximal. Let  $x$  be the distance between  $A$  and the point  $P$  on the ceiling vertically above the weight and the pulley. By the Pythagorean theorem, the distance between the pulley and the ceiling is  $\sqrt{r^2 - x^2}$ . Similarly, the distance between the weight and the pulley is  $\ell - \sqrt{(d-x)^2 + r^2 - x^2}$ . Hence if  $f(x)$  is the distance between the weight and the ceiling, then

$$f(x) = \sqrt{r^2 - x^2} + \ell - \sqrt{(d-x)^2 + r^2 - x^2}.$$

The state of the system is fully determined by the value  $x \in [0, r]$ . Therefore, it suffices to find the global maximum of  $f$  on the interval  $[0, r]$ . The first derivative of  $f$  is given by

$$f'(x) = \frac{-x}{\sqrt{r^2 - x^2}} - \frac{-(d-x) - x}{\sqrt{(d-x)^2 + r^2 - x^2}} = \frac{-x}{\sqrt{r^2 - x^2}} + \frac{d}{\sqrt{(d-x)^2 + r^2 - x^2}}, \quad x \in (0, r).$$

Let us use once more Sage to confirm this expression:

```
var("r, d, l"); f(x)=sqrt(r^2-x^2)+l-sqrt((d-x)^2+r^2-x^2)
df1(x)=diff(f, x); show(df1(x))
```

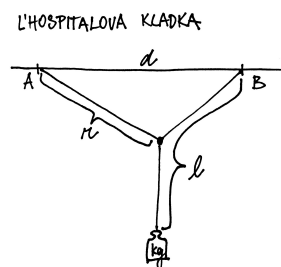
Now, instead of solving the equation  $f'(x) = 0$ , it is easier to solve the equation  $(f'(x))^2 = 0$ , which is equivalent to the polynomial

$$2dx^3 - (2d^2 + r^2)x^2 + d^2r^2 = 0, \quad x \in (0, r).$$

This has  $x = d$  as an obvious solution and hence we get the factorization  $(x - d)(2dx^2 - r^2x - dr^2) = 0$ . Thus, with a bit of try we see that the equation  $f'(x) = 0$  has three solutions. The solution  $x = d$  does not lie on the interval  $[0, r]$  since  $d > r$ . The solution  $x = \frac{r^2 - r\sqrt{r^2 + 8d^2}}{4d}$  is also outside the interval  $[0, r]$ , since  $x_1 < 0$ . But the solution

$$x_0 = \frac{r^2 + r\sqrt{r^2 + 8d^2}}{4d}$$

is positive and since  $r < d$  it satisfies  $x_0 < r$ , i.e.,  $x_0 < \frac{r}{4} + \frac{3r}{4} = r$ . Now, because  $f'$  is continuous on the open interval  $(0, r)$  it changes sign only at  $x_0$ . Based on the limits  $\lim_{x \rightarrow 0^+} f'(x) = \frac{d}{\sqrt{d^2 + r^2}}$  and  $\lim_{x \rightarrow r^-} f'(x) = -\infty$ , we deduce that  $f'(x) > 0$  for all  $x \in (0, x_0)$ , while  $f'(x) < 0$  for all  $x \in (x_0, r)$ . This implies that at  $x_0$  the function  $f$  attains its global maximum.  $\square$



**5.E.126.** A rectangle is inscribed into an equilateral triangle with sides of length  $a$  so that one of its sides lies on one of the triangle's sides and the other two of the rectangle's vertices lie on the remaining sides of the triangle. What is the maximum possible area of the rectangle?

**5.E.127.** Determine the dimensions of an (open) swimming pool whose volume is  $32 \text{ m}^3$  and whose bottom has the shape of a square, so that one would use the least amount of paint possible to prime its bottom and its walls.

**5.E.128.** Express the number 28 as a sum of two non-negative numbers such that the sum of the first summand squared and the second summand cubed is as small as possible.

**5.E.129.** With the help of the first derivative, find the real number  $a > 0$  for which the sum  $a + 1/a$  is minimal. Next solve this problem without using differential calculus.

D) MATERIAL ON INFINITE SUMS AND POWER SERIES

Recall by Section D that infinite sums may have several beautiful applications. Below we describe an application, strongly related with the compelling notion of fractals, the so called Sierpiński carpet.

**5.E.130. Sierpiński carpet.** The unit square is divided into nine equal squares, and then the middle one is removed. Each of the eight remaining squares is again divided into nine equal sub-squares and then the middle sub-square, corresponding to each of these eight squares, is removed again. Having applied this procedure ad infinitum, determine the area of the resulting figure.



**Solution.** In the first step, a square having the area of  $1/9$  is removed. In the second step, eight squares are removed, each having the area of  $9^{-2}$ , so in total we remove  $8 \cdot 9^{-2}$ . Every further iteration removes eight times more squares than in the previous steps, but the squares are nine times smaller. Thus, the sum of areas of all the removed squares is encoded by the following infinite sum:

$$\frac{1}{9} + \frac{8}{9^2} + \frac{8^2}{9^3} + \dots = \sum_{n=0}^{\infty} \frac{8^n}{9^{n+1}}.$$

Now, to compute the area of the remaining figure (known as the *Sierpiński carpet*) it is sufficient to consider the difference

$$1 - \sum_{n=0}^{\infty} \frac{8^n}{9^{n+1}} = 1 - \frac{1}{9} \sum_{n=0}^{\infty} \left(\frac{8}{9}\right)^n = 1 - \frac{1}{9} \cdot \frac{1}{1 - \frac{8}{9}} = 0,$$

where we used the result from 5.B.16 on geometric series. Thus, the Sierpiński carpet has zero area.

**5.E.131.** Decide whether the following implications hold:

(a) Convergence of series  $\sum_{n=1}^{\infty} a_n, \sum_{n=1}^{\infty} b_n$  implies that the series  $\sum_{n=1}^{\infty} (6a_n - 5b_n)$  converges as well.

(b) If a series  $\sum_{n=0}^{\infty} a_n$  satisfies  $\lim_{n \rightarrow \infty} a_n^2 = 0$ , then it is convergent.

(c) If a series  $\sum_{n=1}^{\infty} a_n^2$  converges, then the series  $\sum_{n=1}^{\infty} \frac{a_n}{n}$  converges absolutely.

**5.E.132.** Prove that the series below diverge to  $+\infty$ :

$$T_1 = \sum_{n=2}^{\infty} \frac{1}{\sqrt[n]{\ln(n)}}, \quad T_2 = \sum_{n=1}^{\infty} \frac{1}{n - \ln(n)}.$$

**Solution.** The two enlisted series consist of non-negative terms only. So the series either converge, or diverge to  $+\infty$ . For  $T_1$  it is useful to recall that  $\lim_{n \rightarrow \infty} \sqrt[n]{\ln(n)} = 1$  (see for example 5.E.18), and hence we also get

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt[n]{\ln(n)}} = \frac{1}{\lim_{n \rightarrow \infty} \sqrt[n]{\ln(n)}} = 1.$$

Thus, by (1) in Theorem 5.1.9 we deduce that  $T_1$  is not convergent, and in particular it diverges to  $+\infty$ . For  $T_2$  we have  $\lim_{n \rightarrow \infty} \frac{1}{n - \ln(n)} = 0$ . However, we see that

$$\sum_{n=1}^{\infty} \frac{1}{n - \ln(n)} \geq \sum_{n=1}^{\infty} \frac{1}{n} = +\infty.$$

Hence  $T_2$  also diverges to  $+\infty$ .

**5.E.133.** Prove that the series

$$S = \sum_{n=0}^{\infty} \arctan\left(\frac{n^2 + 2n + 3\sqrt{n} + 4}{n+1}\right), \quad T = \sum_{n=1}^{\infty} \frac{3^n + 1}{n^3 + n^2 - n}$$

are both divergent.

**Solution.** For the first case we have

$$\lim_{n \rightarrow \infty} \arctan\left(\frac{n^2 + 2n + 3\sqrt{n} + 4}{n+1}\right) = \lim_{n \rightarrow \infty} \arctan\left(\frac{n^2}{n}\right) = \frac{\pi}{2}.$$

We can do this computation in Sage by typing

```
var("n"); lim(arctan((n^2+2*n+3*sqrt(n)+4)/(n+1)), n=oo)
```

For the second one we compute

$$\lim_{n \rightarrow \infty} \frac{3^n + 1}{n^3 + n^2 - n} = \lim_{n \rightarrow \infty} \frac{3^n}{n^3} = +\infty.$$

Thus, the necessary condition  $\lim_{n \rightarrow \infty} a_n = 0$  for the series  $\sum_{n=n_0}^{\infty} a_n$  to converge does not hold, and so none of  $S, T$  converges.  $\square$

**5.E.134.** Compute explicitly the sum  $\sum_{n=0}^{\infty} \frac{1}{(3n+1)(3n+4)}$ .

**Solution.** It suffices to use the form (this is the so-called partial fraction decomposition)

$$\frac{1}{(3n+1)(3n+4)} = \frac{1}{3} \cdot \frac{1}{3n+1} - \frac{1}{3} \cdot \frac{1}{3n+4}, \quad n = 0, 1, 2, \dots$$

This gives

$$\sum_{n=0}^{\infty} \frac{1}{(3n+1)(3n+4)} = \lim_{n \rightarrow \infty} \frac{1}{3} \left(1 - \frac{1}{4} + \frac{1}{4} - \frac{1}{7} + \frac{1}{7} - \frac{1}{10} + \dots + \frac{1}{3n+1} - \frac{1}{3n+4}\right) = \lim_{n \rightarrow \infty} \frac{1}{3} \left(1 - \frac{1}{3n+4}\right) = \frac{1}{3}.$$

$\square$

**5.E.135.** (a) Using the ratio test show that  $S = \sum_{n=1}^{\infty} \frac{(-2)^{n^2}}{n!}$  is not convergent.

(b) Exam the convergence of the series  $T = \sum_{n=1}^{\infty} (-1)^{n+1} \arctan\left(\frac{2}{\sqrt{3n}}\right)$ .  $\circ$

**5.E.136.** Let  $S = \sum_{n=0}^{\infty} f(n)$  be an infinite series for which the limit  $\lim_{n \rightarrow \infty} \sqrt[n]{|f(n)|}$  exists, and equals to  $q$ . Present in Sage a syntax based on the root test to deduce, if possible, for the absolute convergence/divergence of  $S$ . Then apply your program for the following series and state Sage's output.



$$S_1 = \sum_{n=1}^{\infty} 2^n \left(\frac{4}{5}\right)^{n^2}, \quad S_2 = \sum_{n=1}^{\infty} \left(\frac{\ln(n)}{n}\right)^n, \quad S_3 = \sum_{n=1}^{\infty} \left(\frac{2n}{4n^2 + 1}\right)^{\frac{1}{n}}, \quad S_4 = \sum_{n=1}^{\infty} (-1)^{n-1} \left(\frac{3n}{n+1}\right)^n.$$

**Solution.** Here one can use once more the command `def` to creating a subroutine, as we did in 5.D.5. This time we will call it `roottest`. Recall that according to the root test the series  $S$  will converge absolutely if  $q < 1$ , will not converge absolutely for  $q > 1$ , while for  $q = 1$  the root test is inconclusive. Thus one can give the cell

```
var("n")
def roottest(f):
    q=lim(abs(f(n))**(1/n), n=oo)
    if q<1:
        return "converges absolutely"
    elif q==1:
        return "no conclusion"
    else:
        return "does not converge absolutely"
    return
```

Now to test our routine via the given series, one can simply introduce first the sequence inside the series, and then apply the command `roottest(f)`. For instance for the second series, type in your editor

`f(n) = (ln(n)/n)^n; roottest(f)`

Sage informs us that the series converges absolutely. Check yourself that the same is true for the first series (see also 5.D.4). On the other hand, the series  $S_4$  does not converge absolutely, while the root test does not provide a result for  $S_3$ .  $\square$

5.E.137. Prove the convergence of the series  $\sum_{n=1}^{\infty} \frac{3^n + 2^n}{6^n}$ , and find its value.  $\circ$

5.E.138. Calculate the series  $\sum_{n=1}^{\infty} \frac{2n-1}{2^n}$  and  $\sum_{n=0}^{\infty} \frac{n+1}{3^n}$ .  $\circ$

5.E.139. Determine those  $\alpha \in \mathbb{R}$ ,  $\beta \in \mathbb{Z}$  and  $\gamma \in \mathbb{R} \setminus \{0\}$  which establish the following series as convergent:

$$\sum_{n=120}^{\infty} \frac{e^{-\alpha n}}{n}, \quad \sum_{n=240}^{\infty} \frac{\beta^n \cdot n!}{n^n}, \quad \sum_{n=360}^{\infty} \frac{n}{\gamma^n}.$$

5.E.140. Determine whether the series  $\sum_{n=21}^{\infty} (-1)^n \frac{n^8 - 5n^6 + 2n}{2^n}$  converges absolutely, converges conditionally, or does not converge at all.  $\circ$

5.E.141. Find all real numbers  $\alpha \geq 0$  for which the series  $\sum_{n=1}^{\infty} (-1)^n \ln(1 + \alpha^{2n})$  is convergent.  $\circ$

5.E.142. Explain, why the convergence interval of the power series expressing the function  $\frac{1}{1+x^2}$  with the center at origin is one.

**Solution.** Recall by 5.B.16 that the geometric series  $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$  converges for  $|x| < 1$ . Thus  $\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n$  and substituting  $x^2$  instead of  $x$ , we arrive at  $\frac{1}{1+x^2} = \sum_{n=0}^{\infty} (-1)^n x^{2n}$ , again converging for  $|x| < 1$ . At the same time, viewing the same sum over complex  $x$ , the sum explodes towards infinite absolute value if  $x$  approaches the imaginary unit  $i$ . Thus the power series cannot converge for  $|x| > 1$ .  $\square$

5.E.143. Express the function  $y = e^x$ , defined on the whole real line, as an infinite polynomial whose terms are of the form  $a_n(x-1)^n$ . Then express the function  $y = 2^x$  defined on  $\mathbb{R}$  as an infinite polynomial with terms  $a_n x^n$ .

**Solution.** We know the series for  $e^x$ , thus writing  $e^{x-1} = e^{-1} e^x$  leads to  $e^x = \sum_{n=0}^{\infty} \frac{e}{n!} (x-1)^n$ . The second task is even simpler, since  $2^x = e^{x \ln(2)} = \sum_{n=0}^{\infty} \frac{\ln^n(2)}{n!} x^n$ .  $\square$

5.E.144. Supposing  $|x| < 1$ , determine the series

$$A(x) = \sum_{n=1}^{\infty} \frac{1}{2n-1} x^{2n-1}, \quad B(x) = \sum_{n=1}^{\infty} n^2 x^{n-1}.$$

5.E.145. Give an example of two divergent series  $\sum_{n=1}^{\infty} a_n$ ,  $\sum_{n=1}^{\infty} b_n$  with positive numbers for which the series  $\sum_{n=1}^{\infty} (3a_n - 2b_n)$  converges absolutely.

5.E.146. Determine whether the series  $\sum_{n=1}^{\infty} (-1)^n \frac{(n!)^2}{(2n)!}$  and  $\sum_{n=1}^{\infty} (-1)^n \frac{n^7 - n^4 + n}{n^8 + 2n^6 + n}$  converge absolutely, converge conditionally, or do not converge at all.

5.E.147. Does the series  $\sum_{n=1}^{\infty} (-1)^{n+1} \frac{\sqrt[3]{n} + \sqrt[5]{n} + 1}{n + \sqrt[5]{n}}$  converge?

5.E.148. Find the values of the parameter  $p \in \mathbb{R}$  for which the series  $\sum_{n=1}^{\infty} (-1)^n \sin^n \frac{p}{n}$  converges.

5.E.149. Determine whether the series  $\sum_{n=0}^{\infty} \frac{2^n + (-2)^n}{5^n}$  converges.

5.E.150. Using the power series  $F(x) = \sum_{n=0}^{\infty} (-1)^n (2n + 1)x^{2n}$  with  $x \in (-1, 1)$ , calculate the infinite sum  $S =$

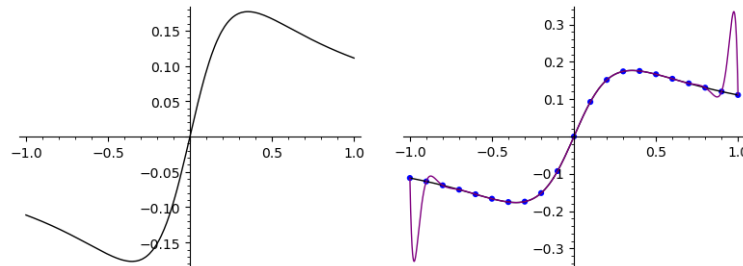
$\sum_{n=1}^{\infty} \frac{2n - 1}{(-2)^{n-1}}$ .

## Solutions to the exercises

**5.A.7.** As usual, we have a taste for Sage, where an appropriate cell reads as follows:

```
g(x)=(x/(1+8*x**2))
figg=plot(g(x), (x, -1, 1), figsize=4, color="black"); show(figg)
R=PolynomialRing(QQ, "x")
n=20
points=[(-1+k*(2/n), g(-1+k*2/n)) for k in [0, 1, ..., n]]
P(x)=R.lagrange_polynomial(points); show(P(x))
figg+=plot(P(x), (x, -1, 1), color="purple")
figg+=list_plot(points, size=20, figsize=4, color="blue"); figg
```

This returns the interpolation polynomial  $P(x)$  under question, and also the plots of  $g(x)$  and of  $P(x)$ , along with the given set of points.



Note that the polynomial oscillates wildly near the endpoints, getting worse and worse as  $n$  increases.

**5.A.8.** An implementation of the following cell in Sage prints out the explicit for of  $P$ , given by  $P(x) = 3x^2 - 2x - 4$ .

```
R = PolynomialRing(QQbar, "x")
R.lagrange_polynomial([( -1, 1), (1, -3), (2, 4)])
```

For further practicing with the method of elementary Lagrange polynomials, we suggest to the reader to verify this answer also by a formal computation.

**5.A.9.** Assume that that  $P(x) = ax^3 + bx^2 + cx + d$  for some reals  $a, b, c, d$  to be specified. By the given conditions one gets the following system (we present this directly in Sage)

```
var("a, b, c, d")
eq1=d-1; eq2=a+b+c+d
eq3=8*a+4*b+2*c+d-1; eq4=27*a+9*b+3*c+d-10
solve([eq1==0, eq2==0, eq3==0, eq4==0], a, b, c, d)
```

This blocks returns the answer  $a = 1, b = -2, c = 0, d = 1$ , that is,  $P(x) = x^3 - 2x^2 + 1$ . Alternatively, one may apply the previous method of Lagrange polynomials:

```
R = PolynomialRing(QQbar, "x")
R.lagrange_polynomial([(0, 1), (1, 0), (2, 1), (3, 10)])
```

which verifies the previous answer:  $x^3 - 2 * x^2 + 1$ .

**5.A.13.** The polynomial  $Q$  is given by  $Q(x) = x^3 - 2x^2 + 3x - 3$ .

**5.A.14.** An answer is given by  $x^5 - 2x^4 - 5x + 2$ .

**5.A.17.** Sought spline differs from the one in 5.A.15 only in the values of the derivatives at the points  $-1$  and  $1$ . Similarly to the previous task, we get that the parts  $S_1$  and  $S_2$  of our spline have the forms  $S_1(x) = ax^3 + bx^2 + 1$  and  $S_2(x) = -ax^3 + bx^2 + 1$ , respectively, where  $a, b, c, d$  are unknown real parameters. Confronting this with the conditions  $S_1(-1) = 0, S_1'(-1) = 1, S_2(1) = 0$ , and  $S_2'(1) = 1$  yields the system  $\{-a + b + 1 = 0, 3a - 2b = -1\}$ , having the solution  $a = -3, b = -4$ . Hence,

$$S(x) = \begin{cases} -3x^3 - 4x^2 + 1, & \text{if } x \in [-1, 0], \\ 3x^3 - 4x^2 + 1, & \text{if } x \in [0, 1]. \end{cases}$$

**5.A.20.**  $S_1(x) = 1 - \frac{11}{20}x + \frac{1}{20}x^3; S_2(x) = \frac{1}{2} - \frac{2}{5}(x-1) + \frac{3}{20}(x-1)^2 - \frac{1}{40}(x-1)^3$ .

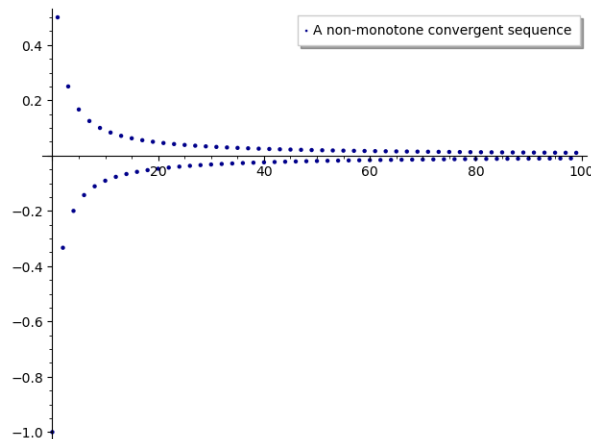
**5.B.1.** The set  $\mathbb{Z}_+$  is unbounded above, hence  $A$  is also unbounded and in particular  $\sup A$  cannot exist. Observe also that  $\left| \frac{(-1)^n}{n} \right| = \frac{1}{n} \leq 1$  and thus we have the inequalities  $n + \frac{(-1)^n}{n} \geq n - \left| \frac{(-1)^n}{n} \right| \geq n - 1$  for all  $n \in \mathbb{Z}_+$ . This implies that  $n + \frac{(-1)^n}{n} \geq 0$  for all  $n \in \mathbb{Z}_+$ , and hence 0 is a lower bound of  $A$ . In particular,  $0 \in A$  and  $A$  is non-empty, thus we deduce that  $\inf A = 0$ .

**5.B.2.** For the set  $B$  we get  $\sup B = \frac{1}{4} \in B$  and  $\inf B = -1 \in B$ . For  $C$  we see that  $\sup C = 9 \notin C$ ,  $\inf C = -9 \notin C$ . For the set  $X$ ,  $\inf X = -1 \in X$ , while  $\sup X = 0 \notin X$ . Further,  $\inf Y = 0 \notin Y$ , while  $\sup Y = 5 \in Y$ .

**5.B.6.** (a) By 5.B.3 we know that  $\frac{1}{n} \rightarrow 0$  and  $3^n \rightarrow +\infty$ . Thus their sum will diverge to  $+\infty$ .  
 (b) Here we are based on fact that if  $a_n \rightarrow +\infty$  and  $b_n \rightarrow +\infty$ , then also  $a_n + b_n = +\infty$ . We have  $n^2 \rightarrow +\infty$  and  $2^n \rightarrow +\infty$ , so  $(n^2 + 2^n) \rightarrow +\infty$ , as well.  
 (c) For this use the first assertion in Problem 5.B.5: Since  $n^2 \rightarrow +\infty$ , and  $3^n \rightarrow +\infty$ , also their product  $n^2 \cdot 3^n \rightarrow +\infty$ .  
 (d) For this use the second assertion in Problem 5.B.5: Since  $\frac{n}{n+1} \rightarrow 1 > 0$  and  $4^n \rightarrow +\infty$ , their product will tend to  $+\infty$  as well.

**5.B.7.** (a) For any  $\varepsilon > 0$  the inequality  $1/2^n < \varepsilon$  is equivalent to  $2^n > 1/\varepsilon$ . Thus, taking  $N > 1/\varepsilon$  we see that  $n \geq N$  implies  $2^n \geq 2^N > N > 1/\varepsilon$ . Hence  $1/2^n < \varepsilon$ , and this proves the claim.  
 (b) Observe that the terms of  $(b_n = (-1)^{n+1})$  interchange between 1 and  $-1$ . Such a sequence is said to be an *alternating sequence*, and since its  $n$ th term is either 1 or  $-1$ , it must diverge by oscillation. For instance, taking  $\varepsilon$  with  $0 < \varepsilon < 1/2$  we see that infinitely many terms of  $(b_n)$  do *not* sit in the open interval  $(x - \varepsilon, x + \varepsilon)$ , for any  $x \in \mathbb{R}$ . Hence the inequality  $|b_n - x| < \varepsilon$  will not make sense for infinitely many  $n$ .  
 (c) The sequence  $(c_n = \sin(n\pi/2))$  ( $n \geq 1$ ) is of the form  $(1, 0, -1, 0, 1, 0, \dots)$ . Therefore, by the same reasoning as in the previous case (b) one deduces that  $(c_n)$  diverges.  
 (d) In this case a solution is given by  $x_n = n, y_n = -n + 1$ , with  $n \in \mathbb{Z}_+$ .

**5.B.12.** We have to prove the *incorrectness* of the statement: “Any convergent sequence is bounded and monotone.” Take for example the sequence  $((-1)^n/n)_{n \geq 1}$ . It converges to 0 but is not monotone, as we can see in the figure below.



For the second question, an example is given by  $(a_n = n)_{n \geq 1}$ . Clearly, this sequence is strictly increasing, but is not bounded and so it cannot be convergent (it is not hard to prove that any convergent sequence is bounded).

**5.B.13.** Recall that  $|\sin(x)| \leq 1$  for any  $x \in \mathbb{R}$ . Thus  $|a_n| = \left| \frac{1}{n} \sin(n) \right| \leq \left| \frac{1}{n} \right| = \frac{1}{n}$ , which implies that  $|a_n| \leq 1/n \leq 1$ , for any natural  $n \geq 1$ . This is equivalent to say that  $-1 \leq a_n \leq 1$ , and so the sequence  $(a_n)$  is bounded. Next, it is easy to see that  $(b_n = n^2 + 1)$  is bounded below,  $b_n \geq 2$  for any  $n$  (since  $n^2 \geq 1$  for any natural  $n \geq 1$ ). Suppose that  $(b_n)$  is also bounded above. Then there exists  $M \in \mathbb{R}$  such that  $b_n \leq M$  for all  $n$ . This gives  $n^2 \leq M - 1$  or  $n \leq \sqrt{M - 1}$  for all  $n$ , a contradiction. Finally, for  $c_n$  one can verify that  $0 \leq c_n \leq 2$ , which we leave as an exercise (we recommend to plot the first terms of  $(c_n)$ , e.g., via Sage).

**5.B.14.** The sequence  $(a_n = \frac{1}{n} \sin(n))$  is bounded but not monotone. However, it is convergent. Indeed, we have  $-(1/n) \leq a_n \leq (1/n)$ , for all  $n \in \mathbb{Z}_+$ , so using the squeeze theorem we obtain  $a_n \rightarrow 0$ . We can verify this in Sage and also plot some terms of  $a_n$ , as follows:

```
a(n)=sin(n)/n; show(lim(a(n), n=oo))
points([n, a(n)] for n in range(1,200)], color="black")
```

Check yourself Sage’s output. Clearly, the sequence  $b_n$  diverges to  $+\infty$ , so it is not convergent. Also, it is easy to see that the sequence  $c_n = ((-1)^n + n)/n$  behaves as  $(a_n)$ . This means that  $(c_n)$  is bounded, and also convergent, though not monotone. We present the verification only via Sage, together with a syntax giving the plot of its first thirty terms. However, once more we leave for practice the execution of this block.

```
c(n) = ((-1)^n + n)/n; show(limit(c(n), n=oo))
list_plot([c(n) for n in range(1, 30)], color="black")
```

Next, one can prove that  $f_n = \frac{1}{\sqrt{n}} = n^{-1/2}$  is decreasing: Since  $\sqrt{n} \leq \sqrt{n+1}$  we have  $\frac{1}{\sqrt{n+1}} \leq \frac{1}{\sqrt{n}}$ , for all  $n \in \mathbb{Z}_+$ . It is also bounded: For any  $n \in \mathbb{Z}_+$  we have  $n \geq 1$ , so  $\sqrt{n} \geq 1$  and thus  $1/\sqrt{n} \leq 1$ . Therefore, by the monotone sequence theorem  $f_n$  should be convergent, and in particular we see that

$$\lim_{n \rightarrow +\infty} f_n = \frac{0}{+\infty} = 0 = \sup \left\{ \frac{1}{\sqrt{n}} : n \in \mathbb{Z}_+ \right\}.$$

Similarly for the sequence  $g_n = \frac{n!}{n^n}$ , which is decreasing and bounded. Thus, the monotone sequence theorem ensures its convergence, and in particular we deduce that

$$\lim_{n \rightarrow +\infty} g_n = 0 = \sup \left\{ \frac{n!}{n^n} : n \in \mathbb{Z}_+ \right\}.$$

This also follows by squeeze theorem, since we have  $0 < g_n \leq 1/n$  for all positive naturals  $n$ . This is because  $n! = n \cdot (n-1) \cdots 2 \cdot 1 < n \cdot n \cdots n \cdot 1 = n^{n-1}$  (so dividing by  $1/n^n$  we obtain the formula).

**5.B.18.** By the binomial theorem we have

$$e_n = 1 + \binom{n}{1} \frac{1}{n} + \cdots + \binom{n}{k} \frac{1}{n^k} + \cdots + \binom{n}{n} \frac{1}{n^n} = 2 + \binom{n}{2} \frac{1}{n^2} + \cdots + \binom{n}{k} \frac{1}{n^k} + \cdots + \binom{n}{n} \frac{1}{n^n}.$$

Since  $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!}$ , for the general term we see that

$$\binom{n}{k} \frac{1}{n^k} = \frac{1}{k!} \cdot \frac{n \cdot (n-1) \cdots (n-k+1)}{n^k} = \frac{1}{k!} \cdot \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} = \frac{1}{k!} \cdot \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right).$$

Therefore

$$e_n = 2 + \frac{1}{2!} \left(1 - \frac{1}{n}\right) + \frac{1}{3!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{k!} \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) + \cdots + \frac{1}{n^n}.$$

As  $n$  increases, the quantities  $\frac{1}{n}, \frac{2}{n}, \dots, \frac{k-1}{n}$  decrease and so the expressions  $(1 - \frac{1}{n}), (1 - \frac{2}{n}), \dots, (1 - \frac{k-1}{n})$  increase. So the general term of  $(e_n)$  increases, i.e.,  $(e_n)$  is increasing. Another way to obtain the monotonicity of  $(e_n)$  is based on the geometric mean inequality and in particular in Bernoulli’s inequality which we presented in Chapter 1. Also, since  $k! = 1 \cdot 2 \cdots k \geq 1 \cdot 2 \cdots 2 = 2^{k-1}$  we see that

$$2 < e_n < 1 + \sum_{k=1}^n \frac{1}{k!} < 1 + \sum_{k=1}^n \frac{1}{2^{k-1}} = 1 + \frac{1 - (1/2)^n}{1 - (1/2)} < 3,$$

since  $1 + \frac{1 - (1/2)^n}{1 - (1/2)} < 1 + \frac{1}{1 - (1/2)}$ . Thus  $(e_n)$  is also bounded,  $2 < e_n < 3$ , for any positive natural  $n$ , and by 5.B.11 it is convergent.

**5.B.24.** Yes: The range of  $f(x) = x^2$  is the set  $[0, +\infty)$ , which according to the remark in 5.B.23 is a closed subset of  $\mathbb{R}$ .

**5.B.29.** This is because the sine function  $f(x) = \sin(x)$  is continuous at any  $x \in \mathbb{R}$ . Verify this claim yourself, presenting a “delta-epsilon” proof.<sup>21</sup> Thus  $\lim_{x \rightarrow \pi/3} \sin(x) = \sin(\pi/3) = \sqrt{3}/2$ .

Let us now use Sage and combine the `limit` command with the `bool` function, to verify that the left/right limits coincide with the value of sine function at the limit point  $x_0$ . This idea is encoded by the following cell:

```
bool(sin(pi/3)==limit(sin(x), x=pi/3, dir="left"))
bool(sin(pi/3)==limit(sin(x), x=pi/3, dir="right"))
```

For both these commands Sage’s output is `True`, as it was expected.

<sup>21</sup>Hint: Use the identity  $\sin(x) - \sin(x_0) = 2 \cos(\frac{x+x_0}{2}) \sin(\frac{x-x_0}{2})$  and some classical inequalities related to the sine and cosine function.



**5.B.32.** (a) We may multiply the expression inside the given limit by  $\frac{1+\cos(x)}{1+\cos(x)}$ , so that we can form the difference of two squares in the numerator, that is,

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x} &= \lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x} \cdot \frac{1 + \cos(x)}{1 + \cos(x)} = \lim_{x \rightarrow 0} \frac{1 - \cos^2(x)}{x(1 + \cos(x))} = \lim_{x \rightarrow 0} \frac{\sin^2(x)}{x(1 + \cos(x))} \\ &= \lim_{x \rightarrow 0} \frac{\sin(x)}{x} \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{1 + \cos(x)} = 1 \cdot \frac{0}{2} = 0. \end{aligned}$$

(b) Here we will use the so called trigonometric “half-angle” identity of the sine function, that is,  $2 \sin^2\left(\frac{x}{2}\right) = 1 - \cos(x)$ . We have

$$\lim_{x \rightarrow 0} \frac{1 - \cos x}{x^2 \sin(x^2)} = \lim_{x \rightarrow 0} \frac{2 \sin^2\left(\frac{x}{2}\right)}{x^2 \sin(x^2)} = \lim_{x \rightarrow 0} \frac{\frac{1}{2} \sin^2\left(\frac{x}{2}\right)}{\left(\frac{x}{2}\right)^2 \sin(x^2)} = \frac{1}{2} \left( \lim_{x \rightarrow 0} \frac{\sin\left(\frac{x}{2}\right)}{\frac{x}{2}} \right)^2 \cdot \lim_{x \rightarrow 0} \frac{1}{\sin^2(x^2)} = \frac{1}{2} \cdot \infty = \infty.$$

Note that this calculation must be considered “from the back”. Indeed, since the limits on the right-hand side exist (no matter whether finite or infinite) and the expression  $\frac{1}{2} \cdot \infty$  is meaningful (see the note after theorem 5.2.13), the original limit exists as well. On the other hand, observe that if we split the original limit into the product

$$\lim_{x \rightarrow 0} (1 - \cos x) \cdot \lim_{x \rightarrow 0} \frac{1}{x^2 \sin(x^2)},$$

then we will get the  $0 \cdot \infty$  type, which is an indeterminate form and tells us nothing about existence of the original limit. Finally, you may like to confirm your computation by Sage, which can be easily done by the cell

```
f(x)=(1-cos(x))/(x^2*sin(x^2)); limit(f(x), x=0)
```

**5.B.33.** (a) One can decompose the polynomial in the denominator, and simplify the given expression, i.e.,

$$\lim_{x \rightarrow 2} \frac{x - 2}{\sqrt{x^2 - 4}} = \lim_{x \rightarrow 2} \frac{x - 2}{\sqrt{(x - 2)(x + 2)}} = \lim_{x \rightarrow 2} \frac{\sqrt{x - 2}}{\sqrt{x + 2}} = \frac{0}{4} = 0.$$

(b) Here one can exploit the rule for limits of compositions of functions (see 5.2.20). Thus

$$\lim_{x \rightarrow 0} \frac{\sin(\sin(x))}{x} = \lim_{y \rightarrow 0} \frac{\sin(y)}{y} \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1.$$

(c) We see that  $\lim_{x \rightarrow 0} \frac{\sin^2(x)}{x} = \lim_{x \rightarrow 0} \sin(x) \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 0 \cdot 1 = 0$ . Notice the original limit exists because both the right-hand side limits exist and their product is well-defined.

**5.B.34.** For instance, in the first case we have  $\lim_{x \rightarrow 0^-} f(x) = 1$  but  $\lim_{x \rightarrow 0^+} f(x) = +\infty$ . Hence  $\lim_{x \rightarrow 0} f(x)$  does not exist. Let us also explain the situation for  $k$  and similarly are treated the function  $g$  and  $h$ . For the sign function  $\text{sign} : \mathbb{R} \rightarrow \mathbb{R}$  we have

$$k(x) = \text{sign}(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \\ -1, & \text{if } x < 0. \end{cases}$$

Taking the sequence  $(x_n = 1/n)$  we see that  $\text{sign}(1/n) \rightarrow 1$  for  $n \rightarrow +\infty$ , while for the sequence  $(y_n = -1/n)$  we obtain  $\text{sign}(-1/n) \rightarrow -1$ , as  $n \rightarrow +\infty$ . Since we found two real sequences  $(x_n), (y_n)$  with  $\lim_{n \rightarrow +\infty} x_n = c = \lim_{n \rightarrow +\infty} y_n$  (where  $c = 0$ ), and  $\lim_{n \rightarrow +\infty} k(x_n) \neq \lim_{n \rightarrow +\infty} k(y_n)$  we deduce that the limit  $\lim_{x \rightarrow c} k(x)$  does not exist (notice that  $c = 0 \in \mathbb{R}$  is a limit point, see also Proposition 5.2.15). In Sage recall that we can treat limits of functions really quickly. For instance, for the functions  $h, k$  to verify the statement type

```
k(x)=sign(x); h(x)=abs(sin(x))/sin(x); limit(k(x), x=0); limit(h(x), x=0)
```

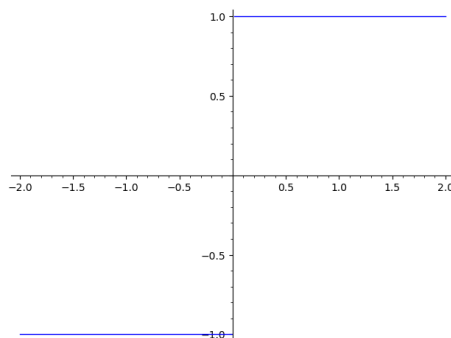
As a side remark notice that the sign function is very similar to the function  $g : \mathbb{R} \setminus \{0\} \rightarrow \{-1, 1\} \cong \mathbb{Z}_2$ , defined by  $g(x) = |x|/x$  which is everywhere continuous except at  $x = 0$ . In particular,  $g(0)$  is not defined and we see that

$$g(x) = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{if } x < 0. \end{cases}$$

If you like to sketch the graph of  $g$ , say for  $-2 \leq x \leq 2$ , type

```
f(x)=abs(x)/x; plot(f, x, -2, 2, exclude=[0])
```

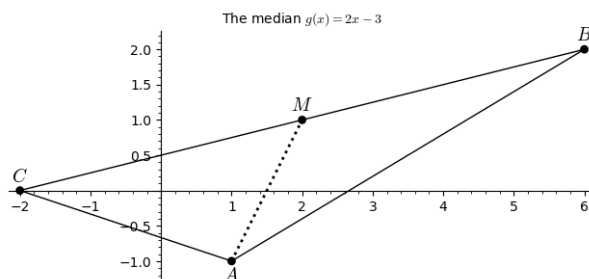
Notice here in our input we added the option `exclude`, to exclude the point that  $g$  is not defined. This syntax produces the known figure, i.e.,



**5.B.36.** Let us multiply both the numerator and denominator by  $1/\alpha^2$ , which yields

$$g(x) = \lim_{\alpha \rightarrow +\infty} \left( \frac{2x - 3 + 4x \frac{1}{\alpha} + \frac{2}{\alpha^2} \cdot \frac{\sin \frac{1}{\alpha}}{\frac{1}{\alpha}}}{1 + \frac{x}{\alpha}} \right) = \lim_{\alpha \rightarrow +\infty} \frac{2x - 3 + 4x \frac{1}{\alpha} + \frac{2}{\alpha^2}}{1 + \frac{x}{\alpha}} \cdot \lim_{\frac{1}{\alpha} \rightarrow 0} \frac{\sin \frac{1}{\alpha}}{\frac{1}{\alpha}} = 2x - 3.$$

This proves that  $g(x) = 2x - 3$ , i.e.,  $g(x)$  represents a line. It remains to prove the second claim. Observe that  $y = 2x - 3$  pass through  $A$ . Let  $M = [x_M, y_M]$  be the middle point of the segment  $BC$ . This has coordinates  $x_M = (6 + (-2))/2 = 2$  and  $y_M = (2 + 0)/2 = 1$ . These coordinates also lie on the line  $y = 2x - 3$ , and hence  $y = g(x)$  is a median of  $ABC$ .



**5.B.37.** (a) We see that

$$\lim_{x \rightarrow +\infty} \frac{3^{x+1} + x^5 - 4x}{3^x + 2^x + x^2} = \lim_{x \rightarrow +\infty} \frac{3 \cdot 3^x}{3^x} = 3.$$

(b) Here one computes

$$\lim_{x \rightarrow +\infty} \frac{4^x - 8x^6 - 2^x - 167}{3^x - 45x - \sqrt{11}\pi^{x+12}} = \lim_{x \rightarrow +\infty} \frac{4^x}{-\sqrt{11}\pi^{12} \cdot \pi^x} = -\infty.$$

(c) In view of the formula  $(a - b)(a + b) = a^2 - b^2$ , we have

$$\lim_{x \rightarrow 0} \frac{\sqrt{1+x} - \sqrt{1-x}}{x} = \lim_{x \rightarrow 0} \frac{(1+x) - (1-x)}{x(\sqrt{1+x} + \sqrt{1-x})} = \lim_{x \rightarrow 0} \frac{2}{\sqrt{1+x} + \sqrt{1-x}} = \frac{2}{\sqrt{1} + \sqrt{1}} = 1.$$

(d) Similarly we see that

$$\begin{aligned} \lim_{x \rightarrow \pi/4} \frac{\cos x - \sin x}{\cos(2x)} &= \lim_{x \rightarrow \pi/4} \frac{(\cos x + \sin x)(\cos x - \sin x)}{(\cos x + \sin x) \cos(2x)} = \lim_{x \rightarrow \pi/4} \frac{\cos^2 x - \sin^2 x}{(\cos x + \sin x) \cos(2x)} \\ &= \lim_{x \rightarrow \pi/4} \frac{1}{\cos x + \sin x} = \frac{1}{\frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}} = \frac{\sqrt{2}}{2}. \end{aligned}$$

Notice the reduction above was made thanks to the identity  $\cos(2x) = \cos^2(x) - \sin^2(x)$ , where  $x \in \mathbb{R}$ .

**5.B.42.** For the function  $f$  we have  $f(x) = x^x = e^{\ln(x^x)} = e^{x \ln(x)}$ . Thus  $f$  is continuous as the composition of the continuous functions  $y_1 = e^x$  and  $y_2 = x \ln(x)$ . Similarly,  $g(x) = x^{\cos(x)} = e^{\cos(x) \ln(x)}$  and  $g$  is continuous as the composition of two continuous functions, namely  $y_1 = e^x$  and  $y_3 = \cos(x) \ln(x)$ .

**5.B.43.** Since  $f, g$  are continuous the given relation means that  $4 \cdot f(2) - g(2) = 2$ , and so  $g(2) = 22$ .

**5.B.44.** We have already seen such an example in 5.B.34. We mean the sign function  $f(x) = \text{sign}(x)$ , which has a ‘‘bump discontinuity’’ at  $x = 0$ . As we have explained in 5.B.34, essentially this is because  $\lim_{x \rightarrow 0^+} f(x) = 1 \neq \lim_{x \rightarrow 0^-} f(x) = -1$ . Sketch the graph of  $f$  to illustrate the situation.

**5.B.46.** The function  $1/x$  is continuous for all  $x \in \mathbb{R} \setminus \{0\}$ , and the function  $\sin(x)$  is continuous for all  $x \in \mathbb{R}$ . Thus the function  $f$  is continuous on the set  $(-\infty, 0) \cup (0, +\infty)$ , i.e., everywhere on  $\mathbb{R}$  except at  $x_0 = 0$ . In particular, at  $x_0 = 0$  it is easy to see that the limit of  $f$  does not exist. The function  $g$  is similar to  $f$ , except of that the oscillations are absorbed by the factor  $x$ . We see that  $-|x| \leq x \sin(\frac{1}{x}) \leq |x|$  and thus by the squeeze theorem we get  $\lim_{x \rightarrow 0} g(x) = 0$ . In particular,  $g$  is continuous everywhere on  $\mathbb{R}$ .

**5.B.49.** The family has the form

$$f_{\alpha,\beta}(x) = \begin{cases} \frac{\sqrt{2x^2 - x + 6} - \alpha x}{x + 2}, & \text{for } x < -2, \\ x^3 + \beta x + 4, & \text{for } x \geq -2, \end{cases}$$

and we see that for all  $x < -2$  the expression  $2x^2 - x + 6$  is positive (and hence  $f_{\alpha,\beta}$  is well-defined). We require the continuity of  $f_{\alpha,\beta}$  at any  $x \in \mathbb{R}$  and hence also at  $x_0 = -2$ , which is equivalent to the condition  $\lim_{x \rightarrow -2^-} f_{\alpha,\beta}(x) = f_{\alpha,\beta}(-2) = \lim_{x \rightarrow -2^+} f_{\alpha,\beta}(x)$ . Then we see that

$$\ell := \lim_{x \rightarrow -2^-} (f_{\alpha,\beta}(x) \cdot (x + 2)) = f_{\alpha,\beta}(-2) \cdot 0 = 0.$$

On the other hand, from the definition of  $f$  and  $\ell$ , we get

$$\ell = \lim_{x \rightarrow -2^-} (\sqrt{2x^2 - x + 6} - \alpha x) = \sqrt{\lim_{x \rightarrow -2^-} (2x^2 - x + 6)} - \lim_{x \rightarrow -2^-} (\alpha x) = 4 + 2\alpha.$$

Thus,  $4 + 2\alpha = 0$ , i.e.,  $\alpha = -2$ . Using this value, we compute

$$\begin{aligned} \lim_{x \rightarrow -2^-} f_{-2,\beta}(x) &= \lim_{x \rightarrow -2^-} \frac{\sqrt{2x^2 - x + 6} + 2x}{x + 2} = \lim_{x \rightarrow -2^-} \frac{2x^2 - x + 6 - 4x^2}{(x + 2)(\sqrt{2x^2 - x + 6} - 2x)} \\ &= \lim_{x \rightarrow -2^-} \frac{-2x^2 - x + 6}{(x + 2)(\sqrt{2x^2 - x + 6} - 2x)} = \lim_{x \rightarrow -2^-} \frac{(x + 2)(-2x + 3)}{(x + 2)(\sqrt{2x^2 - x + 6} - 2x)} \\ &= \lim_{x \rightarrow -2^-} \frac{(-2x + 3)}{(\sqrt{2x^2 - x + 6} - 2x)} = \frac{7}{8}. \end{aligned}$$

On the other hand,  $f_{\alpha,\beta}(-2) = -2(b + 2)$  and we arrive to the equation  $\frac{7}{8} = -2(b + 2)$ , that is,  $b = -\frac{39}{16}$ . Thus only the member  $f_{-2,-\frac{39}{16}}$  is continuous.

Using Sage we can define piecewise functions with many different ways. For instance one can combine the `def` command with `if` and `else`, as follows:

```
def f(x) :
    if x<-2 :
        return((sqrt(2*x^2-x+6)+2*x)/(x+2))
    else :
        return(x^3-(39/16)*x+4)
```

Now, to view  $f$  combine any of the `print` or `show` rules, with the `assume` command. Thus, for example, adding the syntax `assume(x < -2); show(f(x))` will give the first component of  $f$ , and similarly for the second, where instead one should type `assume(x >= -2)`. To evaluate  $f$  at different points  $x \in \mathbb{R}$  just type `f(-4)`, `f(-2)`, `f(0)`, etc. As for the continuity you can add the commands

```
show(bool(limit(f(x), x=-2, dir="+")==f(-2)))
show(bool(limit(f(x), x=-2, dir="-")==f(-2)))
```

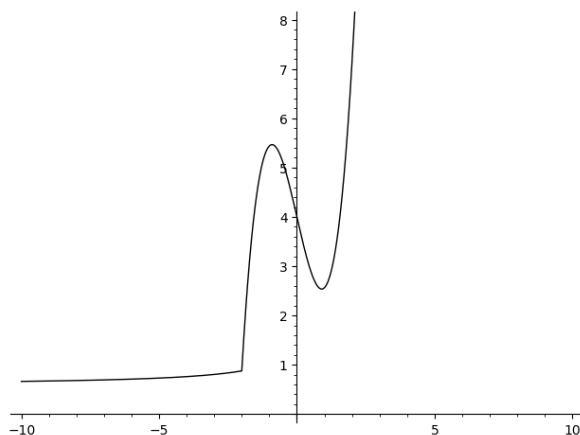
or simply type

```
show(bool(limit(f(x), x=-2)==f(-2)))
```

Finally, to sketch the graph of  $f$  add the cell

```
plot(f, x, -10, 10, ymax=8, color="black")
```

The produced figure is here:



Another way to determine  $f$  and sketch its graph relies on the piecewise function in Sage, which we analyze in 5.C.5. Note that this method currently has some limitations (e.g., in computing limits) and Sage requires a significant amount of time to return an answer (see however in 5.C.5 and 6.D.43 for further details).

**5.B.51.** The given polynomial has at least two roots in  $(-1, 1)$ , This is because  $P(-1) > 0$ ,  $P(0) < 0$ ,  $P(1) > 0$  and thus there must be at least one root in each of these two subintervals, see also 5.2.19.

**5.B.52.** The solution is based on applying Bolzano's theorem via Sage, which goes as follows:

```
f(x) = x^3 - cos(x)*e^x + x*sin(x)
show(f(0)); show(f(pi/2))
```

Sage gives  $f(0) = -1 < 0$  and  $f(1) = \frac{\pi^3}{8} + \frac{\pi}{2} > 0$ , and hence by Bolzano's theorem  $f$  has at least a root in  $(0, \pi/2)$ . As for the implementation of the `find_root` function, just add the syntax `find_root(f, 0, pi/2)`. Sage's answer is 0.9221778114841418.

**5.C.1.** For  $f(x)$ , using the binomial theorem one computes

$$\begin{aligned} (x^n)' &= \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} = \lim_{h \rightarrow 0} \frac{\binom{n}{1}x^{n-1}h + \binom{n}{2}x^{n-2}h^2 + \cdots + h^n}{h} = \\ &= nx^{n-1} + \lim_{h \rightarrow 0} \left( \binom{n}{2}x^{n-2}h + \cdots + h^{n-1} \right) = nx^{n-1}. \end{aligned}$$

For the derivative of  $g(x)$  we rely on identities from trigonometry, as those for  $\sin(x+h)$  and the results from 5.B.31 and 5.B.32, that is,  $\lim_{h \rightarrow 0} \frac{\sin h}{h} = 1$  and  $\lim_{h \rightarrow 0} \frac{\cos h - 1}{h} = 0$ , respectively. We have

$$\begin{aligned} (\sin(x))' &= \lim_{h \rightarrow 0} \frac{\sin(x+h) - \sin(x)}{h} = \lim_{h \rightarrow 0} \frac{\sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x)}{h} = \\ &= \cos(x) \lim_{h \rightarrow 0} \frac{\sin(h)}{h} + \sin(x) \lim_{h \rightarrow 0} \frac{(\cos(h) - 1)}{h} = \cos(x) \cdot 1 + \sin(x) \cdot 0 = \cos(x). \end{aligned}$$

Similarly is treated the square root function  $h(x) = \sqrt{x} = x^{1/2}$ , which is left as an exercise. Here one can show that  $h'(x) = \frac{1}{2\sqrt{x}}$  for all  $x \neq 0$ , but  $f$  is not differentiable at  $x = 0$  (verify this claim by computing the left and right derivatives).

Note that the derivative of the first given function generalizes as  $(x^a)' = ax^{a-1}$ , for all  $x \in \mathbb{R}$  and all  $a > 0$ . When  $a$  is negative, we can use the relation  $(x^a)' = ax^{a-1}$  only for  $x \neq 0$ . These rules provide a way to compute  $h'(x)$  for  $x \neq 0$ :

$$(\sqrt{x})' = (x^{1/2})' = \frac{1}{2}x^{1-1/2} = \frac{1}{2}x^{-1/2} = \frac{1}{2x^{1/2}} = \frac{1}{2\sqrt{x}}.$$

**5.C.2.** The claim is based on the properties of  $f$  and occurs by the definition of the derivative of  $f$ , as follows:

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{f(x)f(h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{f(x)(f(h) - 1)}{h} = f(x) \cdot \lim_{h \rightarrow 0} \frac{f(h) - 1}{h} \\ &= f(x) \cdot \lim_{h \rightarrow 0} \frac{f(0+h) - f(0)}{h} = f(x) \cdot f'(0) = f(x) \cdot 1 = f(x). \end{aligned}$$

**5.C.3.** Recall that

$$f(x) = \begin{cases} x, & \text{if } x \geq 0, \\ -x, & \text{if } x < 0. \end{cases}$$

Thus,  $f'(x) = 1$  for  $x > 0$  and  $f'(x) = -1$  for  $x < 0$ . For instance, for  $x > 0$  we can choose  $h$  small enough such that  $x + h > 0$ , and then

$$f'(x) = \lim_{h \rightarrow 0} \frac{|x+h| - |x|}{h} = \lim_{h \rightarrow 0} \frac{x+h-x}{h} = 1.$$

Similarly for  $x < 0$ . At  $x_0 = 0$  the function  $f$  cannot be differentiable, since the left-side derivative does not agree with the right-side derivative:

$$\lim_{x \rightarrow 0^+} \frac{f(x) - f(0)}{x - 0} = \lim_{x \rightarrow 0^+} \frac{x - 0}{x} = 1, \quad \lim_{x \rightarrow 0^-} \frac{f(x) - f(0)}{x - 0} = \lim_{x \rightarrow 0^-} \frac{-x - 0}{x} = -1.$$

Hence  $f'(0)$  does not exist (and this is what we should expect since the graph of  $f$  forms a corner at 0).

**5.C.4.** If  $f$  is differentiable at  $x_0$ , then the limit  $\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$  exists and equals to  $f'(x_0)$ . Thus we also have

$$(x - x_0) \cdot f'(x_0) = (x - x_0) \cdot \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Consider now the limits in both sides as  $x$  tends to  $x_0$ . In the l.h.s we obtain a zero, since  $\lim_{x \rightarrow x_0} ((x - x_0) \cdot f'(x_0)) = f'(x_0) \cdot \lim_{x \rightarrow x_0} (x - x_0) = f'(x_0) \cdot 0 = 0$ . For the limit in the r.h.s we write altogether

$$0 = \lim_{x \rightarrow x_0} (x - x_0) \cdot \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \rightarrow x_0} \left( (x - x_0) \cdot \frac{f(x) - f(x_0)}{x - x_0} \right) = \lim_{x \rightarrow x_0} (f(x) - f(x_0)),$$

which implies that  $\lim_{x \rightarrow x_0} f(x) = f(x_0)$ . Therefore,  $f$  is continuous at  $x_0$ .

**5.C.6.** In order to use the derivative  $\sin'(x) = \cos(x)$  and the chain rule, we first need to use the identity  $\cos(x) = \sin(x + \frac{\pi}{2})$ . This allows us to view  $\cos(x)$  as the composition  $\cos(x) = f(g(x))$ , with  $f(x) = \sin(x)$  and  $g(x) = x + \frac{\pi}{2}$ , respectively. Then, by the chain rule we get

$$(\cos(x))' = g'(x) \cdot f'(g(x)) = 1 \cdot \cos(x + \frac{\pi}{2}) = \cos(x) \cos(\pi/2) - \sin(x) \sin(\pi/2) = 0 - \sin(x) = -\sin(x).$$

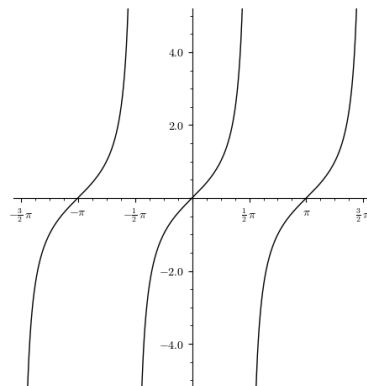
**5.C.10.** We have  $f(1) = \alpha$  and since  $f'(x) = 2\alpha x - 4 \ln(x) - 4$ , it follows that  $f'(1) = 2\alpha - 4$ . Thus the tangent of  $f$  at the point  $P = [1, \alpha]$  is given by

$$y - \alpha = (2\alpha - 4)(x - 1) \iff y = (2\alpha - 4)x - \alpha + 4.$$

We finally see that this line passes through  $[0, 0]$  if and only if  $0 = (2\alpha - 4) \cdot 0 - \alpha + 4$ , that is,  $\alpha = 4$ .

**5.C.13.** The answer is  $y = 2 - x$ ;  $y = x$ .

**5.C.16.** It is useful to recall that the tangent function  $\tan(x) = \sin(x)/\cos(x)$  is not defined at any  $x = \frac{\pi}{2} + \kappa\pi$ , with  $\kappa \in \mathbb{Z}$ . Let us first plot  $\tan(x)$  (in fact, let us present the graph of a periodic extension of the tangent function on  $(-3\pi/2, 3\pi/2)$ ).



To obtain this we used Sage via the following block:

```
p=plot(tan(x), x, -pi/2, pi/2, ymax=5, ymin=-5, color="black")
p+=plot(tan(x-pi), x, -3*pi/2, -pi/2, ymax=5, ymin=-5, color="black",
detect_poles="False")
p+=plot(tan(x+pi), x, pi/2, 3*pi/2, ymax=5, ymin=-5, color="black",
detect_poles="False")
p.show(ticks=pi/2, tick_formatter=pi, aspect_ratio="1")
```

In this block ensure to include the options  $y_{\max}=5$ ,  $y_{\min}=-5$  for plotting correctly the tangent function. From the figure, it is evident that  $\tan(x)$  is strictly increasing for all  $x \in (-\pi/2, \pi/2)$ . Let us substantiate this claim rigorously with a mathematical proof.

Let  $x_1, x_2$  be real numbers such that  $-\frac{\pi}{2} < x_1 < x_2 < \frac{\pi}{2}$ . Based on the identity  $\sin(\alpha - \beta) = \sin(\alpha)\cos(\beta) - \sin(\beta)\cos(\alpha)$  we see that

$$\tan(x_2) - \tan(x_1) = \frac{\sin(x_2)}{\cos(x_2)} - \frac{\sin(x_1)}{\cos(x_1)} = \frac{\sin(x_2)\cos(x_1) - \sin(x_1)\cos(x_2)}{\cos(x_1)\cos(x_2)} = \frac{\sin(x_2 - x_1)}{\cos(x_1)\cos(x_2)} > 0,$$

since  $0 < x_2 - x_1 < \pi$ . Thus, for  $x_1 < x_2$  we have

$$\tan(x_2) - \tan(x_1) > 0 \iff \tan(x_1) < \tan(x_2) \iff f(x_1) < f(x_2),$$

which shows that the tangent function is increasing on  $(-\pi/2, \pi/2)$ . Alternatively, we can compute the first derivative of  $f$ , given by  $f'(x) = \tan'(x) = \frac{1}{\cos^2(x)}$  (see 5.E.76). Thus for any  $x \in (-\pi/2, \pi/2)$  we get  $f'(x) > 0$  and our claim follows.

**5.C.19.** The answer is given by  $\frac{\pi}{6} - \frac{2}{\sqrt{3}} 0.003$ .

**5.C.20.** The answers are  $\frac{1}{2} - \frac{\sqrt{3}\pi}{360}$  and  $\frac{\sqrt{2}}{2} + \frac{\sqrt{2}\pi}{360}$ , respectively.

**5.C.24.** The function  $f$  is continuous on its domain  $[-2, 2]$ . We have  $f'(x) = \frac{2}{3}x^{-\frac{1}{3}} = \frac{2}{3x^{1/3}}$ . Thus  $f'$  is not defined at  $x_0 = 0$ . Since  $x_0 \in (-2, 2)$ , this means that we cannot apply the mean value theorem. In the opposite, for the functions  $g, h$  the mean value theorem applies. Indeed,  $g$  is continuous on  $[1, 4]$  with  $g'(x) = \frac{1}{(x+1)^2}$ . Thus  $g$  is also differentiable on the open interval  $(1, 4)$  and according to the Theorem 5.3.9 there exists  $c \in (1, 4)$  satisfying  $g'(c) = \frac{g(4)-g(1)}{4-1} = \frac{4/5-1/2}{3} = \frac{1}{10}$ . From this we get  $(1+c)^2 = 10$ , that is,  $c = \pm\sqrt{10}-1$  and since  $c \in (1, 4)$  we can accept only the value  $c = \sqrt{10}-1 \approx 2.163$ . Similarly is treated the case for  $h$ .

**5.C.26.** It is sufficient to apply Cauchy's mean value theorem for the functions  $f(x) = x^{n+1}$  and  $g(x) = x^n$ , which are both continuous on  $[a, b]$ , and differentiable on  $(a, b)$ . Also, by assumption  $0 \notin (a, b)$  and hence  $g'(x) = nx^{n-1} \neq 0$  for all  $x \in (a, b)$ . Thus  $f, g$  satisfy the requirements of the Cauchy's mean value theorem. This gives that  $g(b) \neq g(a)$  and moreover there exists  $c \in (a, b)$  such that

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)} \iff \frac{b^{n+1} - a^{n+1}}{b^n - a^n} = \frac{(n+1)c^n}{nc^{n-1}} \iff \frac{n}{n+1} \left( \frac{b^{n+1} - a^{n+1}}{b^n - a^n} \right) = c,$$

for any  $n \in \mathbb{N} \setminus \{0\}$ . Since  $c \in (a, b)$  we are done.

**5.C.31.** (a)  $v(0) = 6$  m/s; (b)  $t = 3$  s,  $s(3) = 16$  m; (c)  $v(4) = -2$  m/s,  $a(4) = -2$  m/s<sup>2</sup>.

**5.C.34.** Such an example occurs by considering the limit  $\lim_{x \rightarrow 0} \frac{\sin(x)}{x+1}$ , which obviously equals to 0 and is not of some indeterminate form. Another example is the limit  $\lim_{x \rightarrow 0} \frac{\sin(x)}{e^x}$ .

**5.D.2.** Often, we may use Sage to evaluate convergent infinite series, or verify summation identities. Recall from Chapter 1 that this can be done via the command `sum`, whose general form is `sum(f(n), n, a, b)` (formally, this corresponds to the sum  $\sum_{n=a}^b f(n)$ ). Of course, one can replace `b` by `Infinity` or its alias `oo`, and this corresponds to infinite sums. This method can be applied for example when one tries to evaluate the zeta function  $\zeta(p)$ , for certain  $p$ , as in our example. In particular, an explicit implementation of our task goes as follows:

```
var("n")
show(sum(1/n^2, n, 1, oo))
show(sum(1/n^3, n, 1, oo))
show(sum(1/n^4, n, 1, oo))
show(sum(1/n^5, n, 1, oo))
show(sum(1/n^6, n, 1, oo))
show(sum(1/n^7, n, 1, oo))
show(sum(1/n^8, n, 1, oo))
show(sum(1/n^9, n, 1, oo))
```

Let us put Sage's output in a table:

$p$	$\zeta(p)$	$p$	$\zeta(p)$
2	$\pi^2/6$	3	$\zeta(3)$
4	$\pi^4/90$	5	$\zeta(5)$
6	$\pi^6/945$	7	$\zeta(7)$
8	$\pi^8/9450$	9	$\zeta(9)$

Hence Sage is able to evaluate explicitly  $\zeta(p)$  for  $p = 2, 4, 6, 8$ . However, as we expected, Sage is unable to compute the cases  $p = 3, 5, 7, 9$ . In these instances, Sage indicates that the sum we are trying to evaluate corresponds to the zeta function.

We will encounter further related applications in the sequel (see for example 5.D.8).

**5.D.7.** The series  $S_1$  converges absolutely. For instance, we see that

$$\sum_{n=1}^{\infty} \left| \frac{\sin(n)}{n^2} \right| \leq \sum_{n=1}^{\infty} \frac{1}{n^2} < \sum_{n=0}^{\infty} \frac{1}{2^n} = 2.$$

Passing to  $S_2$ , one observes that this is an alternating series (since  $\cos(\pi n) = (-1)^n$ ,  $n \in \mathbb{N}$ ). Moreover, we see that the sequence of the absolute values of its terms is decreasing, and  $\lim_{n \rightarrow \infty} \frac{1}{\sqrt[3]{n^2}} = 0$ . It follows that the series  $S_2$  is convergent. In addition, we see that

$$\sum_{n=1}^{\infty} \left| \frac{\cos(\pi n)}{\sqrt[3]{n^2}} \right| = \sum_{n=1}^{\infty} \frac{1}{\sqrt[3]{n^2}} \geq \sum_{n=1}^{\infty} \frac{1}{n} = +\infty,$$

which implies that  $S_2$  converges also conditionally.

**5.D.9.** We see that

$$1 \leq 1, \quad \frac{1}{2^2} + \frac{1}{3^2} < 2 \cdot \frac{1}{2^2} = \frac{1}{2}, \quad \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} < 4 \cdot \frac{1}{4^2} = \frac{1}{4}, \quad \dots,$$

and more general

$$\frac{1}{(2^n)^2} + \dots + \frac{1}{(2^{n+1} - 1)^2} < 2^n \cdot \frac{1}{(2^n)^2} = \frac{1}{2^n}, \quad n = 1, 2, \dots$$

Hence, by comparing the terms of both of the series, we get the required inequality. By the way, notice that from this inequality it follows that the series  $\sum_{n=1}^{\infty} \frac{1}{n^2}$  converges absolutely. Eventually, let us specify that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < 2 = \sum_{n=0}^{\infty} \frac{1}{2^n}.$$

**5.D.15.** The radius of convergence equals to  $r = +\infty$ .

**5.D.16.** The radius of convergence equals to 1.

**5.D.17.** The domain of convergence is the closed interval  $[-1, 1]$ .

**5.D.18.** The answer is  $x \in [2 - \frac{1}{3}, 2 + \frac{1}{3}]$ .

**5.D.19.** The series converges for all  $x \in [-\sqrt[3]{2}, \sqrt[3]{2}]$ .

**5.D.20.** The series converges for all  $x \in [-1, 1]$

**5.E.3.** The obvious answer is  $x^2$ .

**5.E.4.** The natural cubic spline interpolating the given data is given by  $S_1(x) \equiv x$ ;  $S_2(x) \equiv x$ .

**5.E.5.** The complete cubic interpolation spline in question is given by  $S_i(x) = x + 3$ ,  $x \in [-3 + i - 1, -3 + i]$ ;  $i \in \{1, 2\}$ .

**5.E.6.** The mentioned values are given by  $\cos^2(\frac{\pi}{4}) = 1/2$ ,  $\cos^2(\frac{\pi}{3}) = 1/4$ ,  $\cos^2(\frac{\pi}{2}) = 0$ . Since the third value is zero, we need to compute only the values of the first two elementary Lagrange polynomials at the given points. Based on the definition of such polynomials and the given data, it follows that

$$\ell_0(1) = \frac{(1 - \frac{\pi}{3})(1 - \frac{\pi}{2})}{(\frac{\pi}{4} - \frac{\pi}{3})(\frac{\pi}{4} - \frac{\pi}{2})} = 8 \frac{(\pi - 3)(\pi - 2)}{\pi^2}, \quad \ell_1(1) = \frac{(1 - \frac{\pi}{4})(1 - \frac{\pi}{2})}{(\frac{\pi}{3} - \frac{\pi}{4})(\frac{\pi}{3} - \frac{\pi}{2})} = -9 \frac{(\pi - 4)(\pi - 2)}{\pi^2}.$$

Thus, we deduce that

$$P(1) = \frac{1}{2} \cdot 8 \frac{(\pi - 3)(\pi - 2)}{\pi^2} - \frac{1}{4} \cdot 9 \frac{(\pi - 4)(\pi - 2)}{\pi^2} + 0 = \frac{(7\pi - 12)(\pi - 2)}{4\pi^2} \approx 0.288913.$$

Note that the actual value is approximately  $\cos^2 1 \approx 0.291927$ .

**5.E.8.** Such a polynomial is given by  $x^4 + 2x^3 - x^2 + x - 2$ .

**5.E.15.** Consider any singleton (one-element set)  $X \subset \mathbb{R}$ .

**5.E.16.** The set  $C$  must be a singleton. Thus, let us choose  $C = \{0\}$ , for example. Now as  $A, B$  we can take the open subsets  $A = (-1, 0)$ , and  $B = (0, 1)$ , respectively.

**5.E.19.** Using the properties of powers one can express the general term of the given sequence as

$$a_n = \left(\sqrt{2} \cdot \sqrt[4]{2} \cdot \sqrt[8]{2} \cdots \sqrt[2^n]{2}\right) = 2^{\frac{1}{2}} \cdot 2^{\frac{1}{4}} \cdot 2^{\frac{1}{8}} \cdots 2^{\frac{1}{2^n}} = 2^{\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n}}.$$

Thus, in combination with the continuity of the exponential function (a property discussed in the paragraph 5.2.22), we get

$$\lim_{n \rightarrow \infty} a_n = 2^{\lim_{n \rightarrow \infty} \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n}\right)} = 2^{\left(\sum_{n=1}^{\infty} \frac{1}{2^n}\right)}.$$

Next we use a known formula for the sum of geometric series:  $\sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n = \frac{1}{2} \cdot \frac{1}{1 - \frac{1}{2}} = 1$ . Thus finally  $\lim_{n \rightarrow \infty} a_n = 2$ .

Notice in Sage we can simply type

```
n=var("n"); b=2**(sum(1/2**n, n, 1, oo)); b
```

which answers 2, and the claim follows.

**5.E.20.** Notice that every natural number  $k \geq 2$  satisfies the relation  $\frac{1}{(k-1)k} = \frac{1}{k-1} - \frac{1}{k}$ . (this identity is a special case of the so called partial fraction decomposition). Therefore, we compute

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(\frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{(n-1) \cdot n}\right) &= \lim_{n \rightarrow \infty} \left(\frac{1}{1} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \frac{1}{4} + \cdots + \frac{1}{n-1} - \frac{1}{n}\right) \\ &= \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) = 1. \end{aligned}$$

Let us mention that this limit determines the sum of one of the so-called “telescoping series” (used by J. Bernoulli already about 300 years ago). In Sage we can verify the result as usual, i.e., by the cell

```
var("i, n"); limit(sum(1/(i*(i+1)), i, 1, n), n=oo)
```

**5.E.21.** We see that

$$\lim_{n \rightarrow \infty} \left(\frac{1}{n^2} + \frac{2}{n^2} + \cdots + \frac{n-2}{n^2} + \frac{n-1}{n^2}\right) = \lim_{n \rightarrow \infty} \left(\frac{1+n-1}{n^2} \cdot \frac{n-1}{2}\right) = \frac{1}{2}.$$

In Sage we can type `limit(sum(i/n**2, i, 1, n-1), n=oo)`, which provides a quick confirmation of our computation.

**5.E.23.** For these tasks one relies on the relation  $\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a$ , where  $a \in \mathbb{R}$ , see 5.B.18. This means that

$$e^{-1} = \frac{1}{e} = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = \lim_{n \rightarrow \infty} \left(\frac{n-1}{n}\right)^n.$$

Thus, the substitution  $m = n - 1$  gives

$$\lim_{n \rightarrow \infty} \left(\frac{n-1}{n}\right)^n = \lim_{m \rightarrow \infty} \left(\frac{m}{m+1}\right)^{m+1} = \lim_{m \rightarrow \infty} \left(\frac{m}{m+1}\right)^m \cdot \lim_{m \rightarrow \infty} \frac{m}{m+1}.$$

Clearly, the second limit is equal to 1 and replacing  $n$  with  $m$  we get the result

$$\frac{1}{e} = \lim_{n \rightarrow \infty} \left(\frac{n}{n+1}\right)^n.$$

In Sage we can easily verify this by typing `n = var("n"); f(n) = (n/(n+1))n; lim(f(n), n=oo)`. For the second limit we see that

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n^2}\right)^n = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n^2}\right)^{\frac{n^2}{n}} = \lim_{n \rightarrow \infty} \left(\left(1 + \frac{1}{n^2}\right)^{n^2}\right)^{\frac{1}{n}} = e^0 = 1.$$

For the third limit we will apply only Sage and leave to the reader the formal details. Hence type

```
n=var("n"); g(n)=(1-1/n)**(n^2); lim(g(n), n=oo)
```

This returns the value 0.

**5.E.27.** (a) We have  $a > 0$ , thus we also get  $0 < na < na + 1$  and  $0 < \frac{1}{na+1} < \frac{1}{na}$ , for all positive  $n$ . This gives

$$\left|\frac{1}{1+na} - 0\right| = \left|\frac{1}{1+na}\right| = \frac{1}{1+na} < \frac{1}{na} = \frac{1}{a} \cdot \frac{1}{n}, \quad \text{for all } n = 1, 2, \dots$$

As the sequence  $(1/n)$  tends to 0 and  $1/a > 0$ , the inequality obtained above allows us to invoke the result from 5.E.25, with  $n \geq N = 1$ . Thus  $\lim_{n \rightarrow \infty} \left(\frac{1}{1+na}\right) = 0$ .

(b) Obviously, the sequence  $(a_n)$  with general term  $a_n = \frac{x^n}{n!}$  satisfies  $a_n \neq 0$  for all  $n$ . Moreover,



$$\ell = \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n \rightarrow \infty} \left| \frac{x^{n+1}}{(n+1)!} \cdot \frac{n!}{x^n} \right| = \lim_{n \rightarrow \infty} \left| \frac{x}{n+1} \right| = 0 < 1.$$

Thus, by the second result in 5.E.26 we get  $\lim_{n \rightarrow \infty} \frac{x^n}{n!} = 0$ , for all  $x \in \mathbb{R}$ .

**5.E.31.** By definition,  $x_1 = 1 > 0$ . Assuming that  $x_n > 0$  for some  $n$ , we get  $x_{n+1} = 1/(4 + x_n) > 0$ . Thus, by induction we see that  $(x_n)$  is a positive sequence, that is,  $x_n > 0$  for all  $n$ . Using this, for all  $n \geq 2$  one computes

$$|x_{n+1} - x_n| = \left| \frac{1}{4 + x_n} - \frac{1}{4 + x_{n-1}} \right| = \left| \frac{x_{n-1} - x_n}{(4 + x_n)(4 + x_{n-1})} \right| = \frac{|x_{n-1} - x_n|}{(4 + x_n)(4 + x_{n-1})} < \frac{1}{16} |x_n - x_{n-1}|,$$

Thus  $(x_n)$  is a contractive sequence and by 5.E.30 it is Cauchy, and hence convergent.

**5.E.32.** Obviously,  $\{a_k : k \geq n + 1\} \subseteq \{a_k : k \leq n\}$  and hence  $m_{n+1} = \sup\{a_k : k \geq n + 1\} \leq \sup\{a_k : k \geq n\} = m_n$ . This shows that the sequence  $(s_n)$  is decreasing. Similarly, the sequence  $(\ell_n)$  is increasing, since  $\ell_n \leq \inf\{a_k : k \geq n + 1\} = \ell_{n+1}$ , see also the discussion in 5.4.6. By assumption,  $(a_n)$  is bounded hence  $(s_n)$  and  $(\ell_n)$  are also bounded, and the result follows by 5.B.11.

**5.E.36.** The naturals  $\mathbb{N}$  form a closed subset of  $\mathbb{R}$  (recall by 5.B.20 that there are no limit points). Another explanation occurs from the fact that any  $b \notin \mathbb{N}$  has got a small neighbourhood disjoint with  $\mathbb{N}$ . On the other hand, the subset  $\mathbb{Q} \subset \mathbb{R}$  is neither closed nor open. Indeed, in 5.B.20 we saw that the set of all limit points of  $\mathbb{Q}$  is the real line  $\mathbb{R}$ , so  $\mathbb{Q}$  cannot be closed. Moreover, we saw that there are no interior points, so  $\mathbb{Q}$  cannot be open.

**5.E.37.** (a) The open subsets are  $\mathbb{R}^*$  and  $\mathbb{R} \setminus \mathbb{Z}$ .

(b) Only the set  $[1, 2] \cup \{5\}$  is closed.

**5.E.47.** We will show that the sequence is strictly increasing and bounded, so from the monotonicity theorem (5.B.11), it should be convergent. Obviously,  $a_1 = \sqrt{2} \leq a_2 = \sqrt{2 + \sqrt{2}}$ . Assume that  $a_k < a_{k+1}$  for some  $k$ . Using the induction hypothesis we see that  $2 + a_k < 2 + a_{k+1}$ , or  $\sqrt{2 + a_k} < \sqrt{2 + a_{k+1}}$ , that is  $a_{k+1} < a_{k+2}$ . Hence,  $a_n < a_{n+1}$  for all  $n \in \mathbb{N}$ , which means that  $(a_n)_{n \in \mathbb{N}}$  is strictly increasing. Notice now that  $a_1 = \sqrt{2} < 2$ . If it is also  $a_{n-1} < 2$ , then we get  $2 + a_{n-1} < 4$ , that is,  $\sqrt{2 + a_{n-1}} < 2$ , or equivalently  $a_n < 2$ . Thus  $(a_n)$  is bounded and by the monotonicity theorem (5.B.11) it is convergent,  $\lim_{n \rightarrow +\infty} a_n = \sup\{a_n : n \in \mathbb{N}\} = 2$ . To verify this final conclusion, let  $x = \lim_{n \rightarrow +\infty} a_n$ . Then taking the limits in both sides of the relation  $a_n = \sqrt{2 + a_{n-1}}$  we see that  $\sqrt{x + 2} = x$ , which has two solutions ( $x = -1, x = 2$ ), and the acceptable is the positive one.

**5.E.49.** We have  $\lim_{x \rightarrow 0} \frac{\sin^2(x)}{x} = \lim_{x \rightarrow 0} \sin(x) \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 0 \cdot 1 = 0$ . Apparently,  $\lim_{x \rightarrow 0} \frac{x}{\sin(x)} = 1^{-1} = 1$  and the limit  $\lim_{x \rightarrow 0} \frac{1}{\sin(x)}$  does not exist. Similarly, using the rule for the limit of a product we see that the limit  $\lim_{x \rightarrow 0} \frac{x}{\sin^2(x)}$  does not exist.

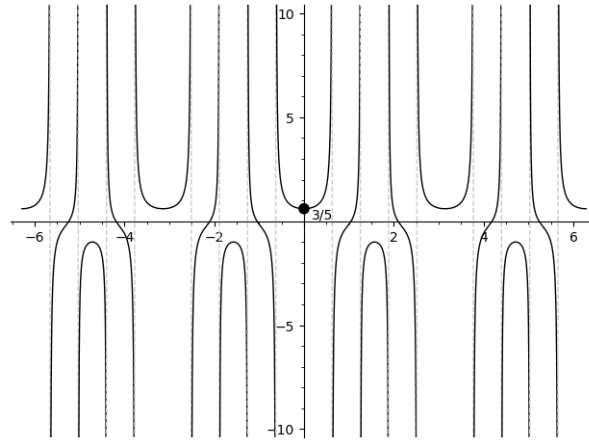
For the calculation of the limit  $\lim_{x \rightarrow 0} \frac{\arcsin(x)}{x}$  use the identity  $x = \sin(\arcsin(x))$ , which makes sense for any  $x \in (-1, 1)$ .

This gives  $\lim_{x \rightarrow 0} \frac{\arcsin(x)}{x} = \lim_{x \rightarrow 0} \frac{\arcsin(x)}{\sin(\arcsin(x))} = \lim_{y \rightarrow 0} \frac{y}{\sin y} = 1$ , where we substitute  $y = \arcsin(x)$ . Observe that  $y \rightarrow 0$  follows from substituting  $x = 0$  into  $y = \arcsin(x)$  and from continuity of this function at 0 (this also guarantees that such a substitution can be made).

Next, see that

$$\lim_{x \rightarrow 0} \frac{3 \tan^2(x)}{5x^2} = \lim_{x \rightarrow 0} \left( \frac{3}{5} \cdot \frac{\sin(x)}{x} \cdot \frac{\sin(x)}{x} \cdot \frac{1}{\cos^2(x)} \right) = \frac{3}{5} \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{x} \cdot \lim_{x \rightarrow 0} \frac{\sin(x)}{x} \cdot \lim_{x \rightarrow 0} \frac{1}{\cos^2(x)} = \frac{3}{5} \cdot 1 \cdot 1 \cdot 1 = \frac{3}{5}.$$

For the next case it is useful to sketch the graph of the function  $\sin(3x)/\sin(5x)$ , via Sage. This is given by



This graph occurs via the following cell in Sage:

```
a=plot(sin(3*x)/sin(5*x), x, -2*pi, 2*pi, detect_poles="show", ymin=-10,ymax=10,
color="black")
b=point((0, 3/5), size = 70, color="black")
c=text("3/5", (0.4, 0.3), color="black", fontsize="10"); show(a+b+c)
```

As we see, the limit under question must be equal to  $3/5$ . Let us verify this in a formal way:

$$\lim_{x \rightarrow 0} \frac{\sin(3x)}{\sin(5x)} = \lim_{x \rightarrow 0} \left( \frac{\sin(3x)}{3x} \cdot \frac{5x}{\sin(5x)} \cdot \frac{3}{5} \right) = \lim_{x \rightarrow 0} \frac{\sin(3x)}{3x} \cdot \lim_{x \rightarrow 0} \frac{5x}{\sin(5x)} \cdot \frac{3}{5} = \lim_{y \rightarrow 0} \frac{\sin y}{y} \cdot \lim_{z \rightarrow 0} \frac{z}{\sin z} \cdot \frac{3}{5} = \frac{3}{5}.$$

For a confirmation via Sage, type `lim(sin(3*x)/sin(5*x), x = 0)`. We leave to the reader to show that the final limit equals to  $3/5$ , as well.

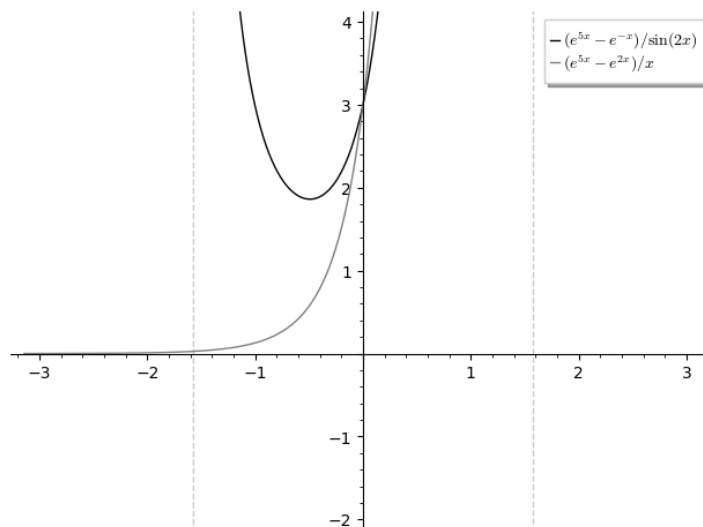
**5.E.50.** Based on the fact that  $\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1$ , for the first limit we obtain

$$\lim_{x \rightarrow 0} \frac{e^{5x} - e^{2x}}{x} = \lim_{x \rightarrow 0} \left( e^{2x} \frac{e^{(5-2)x} - 1}{(5-2)x} (5-2) \right) = \lim_{x \rightarrow 0} e^{2x} \cdot \lim_{x \rightarrow 0} \frac{e^{3x} - 1}{3x} \cdot 3 = e^0 \cdot \lim_{y \rightarrow 0} \frac{e^y - 1}{y} \cdot 3 = 1 \cdot 1 \cdot 3 = 3.$$

In a similar way one computes the second limit, and we leave the details for practice. We present the solution in Sage and include in a figure both plots.

```
G(x) = (e**(5*x) - e**(2*x)) / x; E(x) = (e**(5*x) - e**(-x)) / sin(2*x); show(lim(E(x), x=0))
a=plot(E(x), x, -pi, pi, detect_poles="show", ymin=-2, ymax=4, color="black")
b=plot(G(x), x, -pi, pi, detect_poles="show", ymin=-2, ymax=4, color="grey")
show(a+b)
```

Executing this block we get the confirmation for the limit and desired illustration of the graphs, see here:



**5.E.52.** Obviously,  $\lim_{x \rightarrow +\infty} (2 + \frac{1}{x}) = 2$ ,  $\lim_{x \rightarrow +\infty} \frac{1}{x} = 0$  and  $\lim_{x \rightarrow +\infty} x = +\infty$ . Thus,  $\lim_{x \rightarrow +\infty} (2 + \frac{1}{x})^{\frac{1}{x}} = 2^0 = 1$  and  $\lim_{x \rightarrow +\infty} x^{-x} = \lim_{x \rightarrow +\infty} (\frac{1}{x})^x = 0$ . Equally well, we could first compute the limit  $\lim_{x \rightarrow +\infty} (\frac{1}{x} \cdot \ln(2 + \frac{1}{x})) = 0 \cdot \ln 2 = 0$ , and then, as a generalization of the formula

$$\lim_{x \rightarrow x_0} f(x)^{g(x)} = e^{\lim_{x \rightarrow x_0} (g(x) \ln(f(x)))},$$

we obtain  $\lim_{x \rightarrow +\infty} (2 + \frac{1}{x})^{\frac{1}{x}} = e^{\lim_{x \rightarrow +\infty} (\frac{1}{x} \cdot \ln(2 + \frac{1}{x}))} = e^0 = 1$ . Similarly, we see that  $\lim_{x \rightarrow +\infty} (-x \ln(x)) = -\infty$ , and hence

$$\lim_{x \rightarrow +\infty} x^{-x} = e^{\lim_{x \rightarrow +\infty} (-x \ln(x))} = e^{-\infty} = 0.$$

Finally one must be cautious when calculating the final limit. Both one-sided limits exist, but are different, which implies that this limit does not exist:  $\lim_{x \rightarrow 0^+} e^{\frac{1}{x}} = e^{\lim_{x \rightarrow 0^+} \frac{1}{x}} = e^\infty = \infty$  and  $\lim_{x \rightarrow 0^-} e^{\frac{1}{x}} = e^{\lim_{x \rightarrow 0^-} \frac{1}{x}} = e^{-\infty} = 0$ . In Sage one can directly solve the task by the cell

```
f(x)=(2+(1/x))**(1/x); g(x)=x**(-x); j(x)=e**(1/x)
show([lim(f(x),x=+oo), lim(g(x),x=+oo), lim(j(x),x=0)])
```

**5.E.53.** The first limit equals to  $+\infty$ . For the second one, after multiplying by the fraction  $\frac{\sqrt{x^2+x+x}}{\sqrt{x^2+x+x}}$  it follows that  $\lim_{x \rightarrow +\infty} (\sqrt{x^2+x} - x) = 1/2$ . For the third case we see that  $\lim_{x \rightarrow +\infty} (x \sqrt{1+x^2} - x^2) = 1/2$ . Finally,  $\lim_{x \rightarrow 0^-} \frac{\sqrt{1+\tan(x)} - \sqrt{1-\tan(x)}}{\sin(x)} = 1$ .

**5.E.54.** In Sage we can directly type

```
x, n=var("x, n"); f(x)=((1+2*n*x)**n-(1+n*x)**(2*n))/(x**2); lim(f(x), x=0)
```

and this returns the answer  $-n^3$ . To verify this in a formal way, we will use the binomial theorem. This gives

$$(1 + 2nx)^n = 1 + \binom{n}{1} 2nx + \binom{n}{2} (2nx)^2 + P(x) x^3, \quad (1 + nx)^{2n} = 1 + \binom{2n}{1} nx + \binom{2n}{2} (nx)^2 + Q(x) x^3,$$

for some polynomials  $P, Q$ . Let us emphasize on the fact that this really holds for all  $n \in \mathbb{N}^*$  (if  $n = 1$ , the polynomials  $P, Q$  as zero constants). So, for all  $x \in \mathbb{R}$  we obtain the relations

$$(1 + 2nx)^n = 1 + 2n^2x + 2n^3(n-1)x^2 + P(x)x^3, \quad (1 + nx)^{2n} = 1 + 2n^2x + n^3(2n-1)x^2 + Q(x)x^3.$$

Based on these formulas we are now able to present the result, as follows:

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{(1 + 2nx)^n - (1 + nx)^{2n}}{x^2} &= \lim_{x \rightarrow 0} \frac{(2n^3(n-1) - n^3(2n-1))x^2 + (P(x) - Q(x))x^3}{x^2} \\ &= \lim_{x \rightarrow 0} (-n^3 + (P(x) - Q(x))x) = -n^3 + 0 = -n^3. \end{aligned}$$

**5.E.55.** By the definition of  $(a_n)$  and based on the trigonometric inequality  $|\cos(x)| \leq 1$ , ( $x \in \mathbb{R}$ ), we see that

$$|a_{n+1} - a_n| = \left| \frac{\cos(n+1)}{2^{n+1}} \right| \leq \frac{1}{2^{n+1}} < \frac{1}{2^n}.$$

Thus  $|a_{n+1} - a_n| < \frac{1}{2^n}$  and it follows that  $(a_n)$  is a Cauchy sequence. Thus it converges (in fact one can show that a sequence  $(a_n)$  satisfying  $|a_{n+1} - a_n| < c^n$ , for some  $c$  with  $0 < c < 1$  and for all  $n \in \mathbb{Z}_+$ , is a Cauchy sequence).

**5.E.58.** First, let us calculate the one-sided limits at the point  $x_0 = 0$ :

$$\lim_{x \rightarrow 0^-} (x-1)^{-\operatorname{sgn}(x)} = \lim_{x \rightarrow 0^-} (x-1) = -1, \quad \lim_{x \rightarrow 0^+} (x-1)^{-\operatorname{sgn}(x)} = \lim_{x \rightarrow 0^+} \frac{1}{x-1} = -1,$$

whence  $\lim_{x \rightarrow 0} (x-1)^{-\operatorname{sgn}(x)} = -1$ . However, the function will be continuous at  $0 = 0$ , if and only if  $f(0) = -1$ . Recall by the solution given in 5.B.34 that for the sign function we have adopted the convention  $\operatorname{sgn}(0) = 0$ . Thus  $f(0) = (-1)^0 = 1$ , and the function at hand is not continuous at  $x_0 = 0$ . Similarly, for  $x = 1$  we obtain

$$\lim_{x \rightarrow 1^-} (x-1)^{-\operatorname{sgn}(x)} = \lim_{x \rightarrow 1^-} \frac{1}{x-1} = -\infty, \quad \lim_{x \rightarrow 1^+} (x-1)^{-\operatorname{sgn}(x)} = \lim_{x \rightarrow 1^+} \frac{1}{x-1} = \infty.$$

Hence both one-sided limits at the point 1 exist, yet they differ. Thus the limit  $\lim_{x \rightarrow 1} f(x)$  does not exist, and the function is not continuous at  $x_1 = 1$ , as well.

**5.E.59.** The function is continuous at the points  $-\pi, 0, \pi$ , only right-continuous at the point 2, only left-continuous at the point 3, and discontinuous at 1.

**5.E.60.** The function is continuous if and only if  $p = 2$ .

**5.E.61.** The answer is  $a = 4$ .

**5.E.62.** The given function is continuous at every point of its domain  $\mathbb{R} \setminus \{\pm 1\}$ . Thus, the extended function will be continuous if and only if

$$f(-1) = \lim_{x \rightarrow -1} \left( (x^2 - 1) \sin \frac{2x - 1}{x^2 - 1} \right), \quad f(1) = \lim_{x \rightarrow 1} \left( (x^2 - 1) \sin \frac{2x - 1}{x^2 - 1} \right),$$

such that the limits in the right hand side exist and are finite (if either of these limits did not exist, or are infinite, then  $f$  cannot be extended to a continuous function). Indeed, since any  $x \in \mathbb{R}$ ,  $x \neq \pm 1$  satisfies  $\left| \sin \left( \frac{2x-1}{x^2-1} \right) \right| \leq 1$ , it follows that

$$-|x^2 - 1| \leq f(x) \leq |x^2 - 1|, \quad \forall x \in \mathbb{R}, x \neq \pm 1.$$

Clearly  $\lim_{x \rightarrow \pm 1} |x^2 - 1| = 0$ , and by the squeeze theorem we get  $f(\pm 1) = 0$ .

**5.E.70.** If  $f : [a, b] \rightarrow \mathbb{R}$  is injective then we have  $f(a) \neq f(b)$ , and we may assume that  $f(a) < f(b)$  (otherwise consider  $-f$ ). We will show that  $f$  is strictly increasing. First we will show that for any  $x \in (a, b)$  we have  $f(a) < f(x) < f(b)$ . Assume in the contrary that  $f(x) \leq f(a)$  or  $f(x) \geq f(b)$ . Because  $f$  is injective we essentially assume that  $f(x) < f(a)$  or  $f(x) > f(b)$ .

*Case  $f(x) < f(a)$ :* This means that  $f(x) < f(a) < f(b)$  and by applying (the general version of) the intermediate value theorem for the restriction  $f|_{[x, b]}$  we find some  $x_0 \in (x, b)$  with  $f(x_0) = f(a)$ , a contradiction since  $f$  is injective.

*Case  $f(x) > f(b)$ :* This means that  $f(a) < f(b) < f(x)$  and by applying (the general version of) the intermediate value theorem for the restriction  $f|_{[a, x]}$  we find some  $x_0 \in (a, x)$  with  $f(x_0) = f(b)$ , which is again a contradiction by injectivity.

Based now on our claim and taking  $x_1, x_2 \in \mathbb{R}$  with  $a < x_1 < x_2 < b$  it is easy to see that  $f(x_1) < f(x_2)$ , that is,  $f$  is strictly increasing.

**5.E.77.** We will solve this problem using logarithmic differentiation. In particular, let  $f$  be a positive function for which  $f'(x)$  exists. Then recall that

$$(\ln f(x))' = \frac{f'(x)}{f(x)}, \quad \text{that is,} \quad f'(x) = f(x) \cdot (\ln f(x))'.$$

Now, the given function is differentiable as a product of differentiable functions. Thus, by combining the previous rule with basic properties of the natural logarithm, we obtain

$$\begin{aligned} \left( \frac{\sqrt[4]{x-1} \cdot (x+2)^3}{e^x(x+132)^2} \right)' &= \frac{\sqrt[4]{x-1} \cdot (x+2)^3}{e^x(x+132)^2} \cdot \left[ \ln \frac{\sqrt[4]{x-1} \cdot (x+2)^3}{e^x(x+132)^2} \right]' \\ &= \frac{\sqrt[4]{x-1} \cdot (x+2)^3}{e^x(x+132)^2} \cdot \left[ 3 \ln(x+2) + \frac{1}{4} \ln(x-1) - x \ln(e) - 2 \ln(x+132) \right]' \\ &= \frac{\sqrt[4]{x-1} \cdot (x+2)^3}{e^x(x+132)^2} \left[ \frac{3}{x+2} + \frac{1}{4(x-1)} - 1 - \frac{2}{x+132} \right]. \end{aligned}$$

In Sage to verify this complicated computation use the cell

```
f(x) = ((x-1)**(1/4)) * (x+2)^3 / (exp(x) * (x+132)^2)
show(diff(f(x), x).simplify())
```

**5.E.78.** (1)  $a'(x) = x^2 \sin(x)$ ; (2)  $b'(x) = \cos(\sin(x)) \cos(x)$ ; (3)  $c'(x) = \frac{3x^2+2}{x^3+2x} \cos(\ln(x^3+2x))$ ; (4)  $d'(x) = \frac{2(1-2x)}{(1-x+x^2)^2}$ ; (5)  $\varepsilon'(x) = \frac{7}{8} x^{-\frac{1}{8}}$ ; (6)  $f'(x) = \cos(x) \cos(\sin(x)) \cos(\sin(\sin(x)))$ ; (7)  $g'(x) = \frac{\cos x}{3\sqrt[3]{\sin^2 x}}$ ; (8)  $h'(x) = \frac{2x^2}{1-x^6} \sqrt[3]{\frac{1+x^3}{1-x^3}}$ .

**5.E.79.** (a) Such a function is the floor function  $f(x) = \lfloor x \rfloor$  introduced in 5.B.48. For instance, we have  $\lfloor -2.2 \rfloor = -3$  and  $\lfloor 2.2 \rfloor = 2$ , thus  $f(-x) \neq f(x)$  and  $f$  cannot be even. Moreover,  $\lfloor -2.2 \rfloor = -3 \neq -2 = -\lfloor 2.2 \rfloor$ , and hence the floor function is neither odd.

(b) Suppose that  $f$  is even, that is,  $f(-x) = f(x)$  for all  $x$  in the domain of  $f$ . Taking the derivatives in both sides of this equation and based on the chain rule, we get  $f'(x) = (f(-x))' = (-x)' f'(-x) = -f'(-x)$ . This means that  $f'(-x) = -f'(x)$  and hence  $f'$  is odd. Similarly the case with  $f$  odd.

**5.E.80.** The derivative  $f'(x)$  of a polynomial  $f$  provides the slope of the tangent line to its graph at the point  $[x, f(x)] \in \mathbb{R}^2$ . In degree zero,  $f$  is a non-zero constant  $a \cdot x^0 = a \cdot 1 = a$  and its derivative vanishes everywhere. This confirms the fact that the graph in this case is the line  $y = a$ , parallel to the  $x$ -axis. In degree one,  $f(x) = ax + b$  for some  $a, b \in \mathbb{R}$  with  $a \neq 0$ , thus its derivative is the constant  $a$ . Consequently, the slope of the tangent is  $a$  everywhere, which is the property

characterizing a line. In degree two,  $f(x) = ax^2 + bx + c$  with  $a, b, c \in \mathbb{R}$  and  $a \neq 0$ . Obviously,  $f'(x) = 2ax + b$  and  $x = -\frac{b}{2a}$  is the only point where the slope is zero, i.e., the maximum or the minimum of the quadratic function, as pointed out in 5.A.2. In degree three  $f(x) = ax^3 + bx^2 + cx + d$  with  $a, b, c, d \in \mathbb{R}$  and  $a \neq 0$ . The derivative is the second degree polynomial  $f'(x) = 3ax^2 + 2bx + c$ . Clearly, there are points with tangent slope zero if and only if the latter quadratic expressions allows zero values. This happens if and only if its discriminant  $4(b^2 - 3ac)$  is non-negative. In particular, there will be two bumps on the graph if and only if  $b^2 > 3ac$ . The slope of the graph gets to zero only in one point if  $b^2 = 3ac$ , and then the graph curve looks similarly as the graph of  $f(x) = x^3$  at  $x = 0$ .

**5.E.82.** Observe first that  $(\sinh(x))' > 0$  for all  $x \in \mathbb{R}$ , hence  $f$  is strictly increasing on  $\mathbb{R}$  (as we see also in the graph of  $f$  above). Thus it has an inverse function, called the “inverse hyperbolic sine function” and denoted by  $\sinh^{-1}(x)$ . The domain and the range of  $f$  are both  $(-\infty, \infty)$ , hence the same applies for  $\sinh^{-1}$ . To compute its explicit form, let  $y = \sinh^{-1}(x)$ . Then  $x = \sinh(y) = \frac{e^y - e^{-y}}{2}$ , that is,  $e^y - e^{-y} - 2x = 0$ . Multiplying both sides of this equation by  $e^y$  we arrive to the equation  $e^{2y} - 2xe^y - 1 = 0$ , which is obviously a quadratic equation in the variable  $t = e^y$ . Its solutions have the form  $t_{\pm} = x \pm \sqrt{x^2 + 1}$ . However,  $t = e^y > 0$ , but  $x - \sqrt{x^2 + 1} < 0$ , from where we deduce that  $e^y = x + \sqrt{x^2 + 1}$ , or equivalently  $y = \ln(x + \sqrt{x^2 + 1})$ , i.e.,  $\sinh^{-1}(x) = \ln(x + \sqrt{x^2 + 1})$  for all  $x \in \mathbb{R}$ . You may like to confirm this expression in Sage, based on the command `arcsinh` which corresponds to the function  $\sinh^{-1}$ . This, for example, can be done via the cell

```
bool(arcsinh(x)==ln(x+sqrt(x^2+1)))
```

Recall now that  $\ln(f(x))' = f'(x)/f(x)$  and  $(\sqrt{g(x)})' = g'(x)/2\sqrt{g(x)}$ , where  $f, g$  are both positive on their domain. Thus,

$$(\sinh^{-1}(x))' = \left(\ln(x + \sqrt{x^2 + 1})\right)' = \frac{1 + \frac{x}{\sqrt{x^2 + 1}}}{x + \sqrt{x^2 + 1}} = \frac{1}{\sqrt{x^2 + 1}}, \quad x \in \mathbb{R}.$$

As a confirmation via Sage, just type

```
show(diff(ln(x+sqrt(x^2+1)), x).factor())
```

**5.E.86.** (1) It does not because the one-sided derivatives differ. On particular, we get  $\pi/2$  from the right and  $-\pi/2$  from the left.

(2) It does not.

(3) To provide an answer one should recall the derivative of absolute value function. In particular, the function  $f(x) := |x - 5| + |x - 9|$  has the desired properties.

(4) Suppose that  $f(x) = g(x) = 1$  at the rational numbers and  $f(x) = g(x) = -1$  at the irrational numbers.

(5) It is not hard to see that  $f'(x) < 0$ , for all  $x > e$ .

**5.E.87.** We see that  $f'(x) = 5 - 4\sin(x) > 0$ . This means that  $f$  is increasing and hence a bijection. Thus  $f^{-1}$  exists and since  $f(0) = 4$  we have  $f^{-1}(4) = 0$ . Moreover,  $f'(f^{-1}(4)) = f'(0) = 5 \neq 4$ , hence we may apply 5.3.6 which gives

$$(f^{-1})'(4) = \frac{1}{f'(f^{-1}(4))} = \frac{1}{f'(0)} = \frac{1}{5}.$$

**5.E.89.** From the conditions  $P(0) = 1$  and  $P'(0) = 1$  it follows that  $P(x) = bx^3 + cx^2 + x + 1$ , for some  $b, c \in \mathbb{R}$ . The two remaining conditions determine two equations for the variables  $b$  and  $c$ :  $b+c+2 = 2a+2$ ,  $3b+2c+1 = 5a+1$  with the unique solution  $b = c = a$ . Therefore, a polynomial satisfying the desired conditions should has the form  $P(x) = ax^3 + ax^2 + x + 1$ , with  $a \in \mathbb{R}$ . The derivative of  $P$  is a parabola given by  $P'(x) = 3ax^2 + 2ax + 1$  and we require  $P'(x) > 0$  or  $P'(x) < 0$ , which is equivalent saying that the discriminant  $\Delta = 4a^2 - 12a$  of  $P'$  is negative,  $\Delta < 0$ . This gives  $4a(a - 3) < 0$ , which is true for all  $a \in (0, 3)$ .

**5.E.91.** Let  $x_0 \in \mathbb{R}$  some arbitrary point. By means of the definition of the first derivative of a function on  $\mathbb{R}$ , and in combination with the relation  $f(x + y) = \frac{f(x)+f(y)}{1-f(x)f(y)}$  we see that

$$\begin{aligned} f'(x_0) &= \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{h \rightarrow 0} \frac{\frac{f(x_0)+f(h)}{1-f(x_0)f(h)} - f(x_0)}{h} = \lim_{h \rightarrow 0} \frac{f(x_0) + f(h) - f(x_0) + f^2(x_0)f(h)}{h(1 - f(x_0)f(h))} \\ &= \lim_{h \rightarrow 0} \frac{1 + f^2(x_0)}{1 - f(x_0)f(h)} \cdot \lim_{h \rightarrow 0} \frac{f(h)}{h} = \frac{1 + f^2(x_0)}{1 - f(x_0)f(0)} \cdot f'(0) = f'(0) \cdot (1 + f^2(x_0)). \end{aligned}$$

Here we have used that  $f(0) = 0$ , which we get from the relation  $f(x+y) = \frac{f(x)+f(y)}{1-f(x)f(y)}$  by setting  $y = 0$  and moreover that  $f$  is differentiable at 0, and hence  $\lim_{h \rightarrow 0} \frac{f(h)}{h} = \lim_{h \rightarrow 0} \frac{f(h+0)-f(0)}{h-0} = f'(0)$ . Thus  $f'(x) = f'(0)(1+f^2(x))$  for all  $x \in \mathbb{R}$  and  $f$  is differentiable on  $\mathbb{R}$ .

**5.E.95.** (a) The function has a local maximum at the point  $x_1 = e^{-2}$ . It has a local minimum at the point  $x_2 = 1$ .

(b) The answer is  $\frac{1}{\sqrt[3]{e}}$ .

(c) The answer is  $4 = p(-1) = p(2)$ ,  $-16 = p(-3)$ .

(d) No. For instance, if  $\alpha = \sqrt{2}/2$ , there is only a local extremum at the point.

**5.E.96.** The given function  $f$  has domain the whole real line and we have  $f'(x) = -e^{-x} = -\frac{1}{e^x} < 0$  for any  $x \in \mathbb{R}$ . Thus  $f$  is strictly decreasing on  $\mathbb{R}$ .

**5.E.97.** Obviously, a solution is given by  $x = 1$ . We will show that it is unique by using the function  $f(x) = x^{2025} + 2024x + \ln x - 2025$ . It is sufficient to show that  $f$  is strictly growing throughout its domain  $A = (0, +\infty)$ . Obviously,  $f$  is differentiable over  $A$  and its derivative is given by  $f'(x) = 2024x^{2024} + 2024 + \frac{1}{x}$ . But  $f'(x) > 0$  for all  $x \in A$ , so we are done.

**5.E.103.** Consider the function  $g(x) = x - \ln(x+1)$ , with  $g(0) = 0$ . The first derivative of  $g$  is given by

$$g'(x) = 1 - \frac{1}{x+1} = \frac{x}{x+1}.$$

Thus  $g'(x) = 0$  if and only if  $x = 0$ . For  $x > 0$  we have obviously  $g'(x) > 0$ , so  $g$  is strictly increasing on the open interval  $(0, +\infty)$ . Hence, for all  $x > 0$  we have  $g(x) > g(0) = 0$ , that is,  $x > \ln(x+1)$ . Since for  $x = 0$  this holds as an equality we have finally proved  $x \geq \ln(x+1)$  for all  $x \geq 0$ .

**5.E.104.** Consider the function  $f(x) = \alpha x^\beta - \beta x^\alpha - \alpha + \beta$ . We have  $f'(x) = \alpha\beta(x^{\beta-1} - x^{\alpha-1}) > 0$  for all  $x > 1$ , that is,  $f(x)$  is strictly increasing for all  $x \in (1, +\infty)$ . The point  $x = 1$  induces a critical point for  $f$ , i.e.,  $f'(1) = 0$  and we see that there  $f$  attains its minimum value  $f(1) = 0$ . Thus for all  $x > 1$  we have  $f(x) > 0$ , i.e.,  $\alpha x^\beta - \beta x^\alpha > \alpha - \beta$ .

**5.E.108.** (a) Obviously, this limit has the indeterminate form  $\frac{\infty}{\infty}$ .

(b) The l'Hospital's rule gives rise to the following limit

$$\lim_{x \rightarrow +\infty} \frac{x + \sin(x)}{x} = \lim_{x \rightarrow +\infty} \frac{1 + \cos(x)}{1} = \lim_{x \rightarrow +\infty} (1 + \cos(x)),$$

which does *not* exist (e.g., use Sage and type the syntax `limit(1 + cos(x), x = oo)`). However, the limit at hand exists. This is because

$$\frac{x-1}{x} \leq \frac{x + \sin x}{x} \leq \frac{x+1}{x},$$

which implies that  $\lim_{x \rightarrow +\infty} f(x) = 1$ .

**5.E.111.** The answers are  $a = 2/\pi$ ,  $b = 1$ ,  $c = 1/2$  and  $d = e^{-2}$ .

**5.E.116.** The perimeter is  $2\sqrt{5}r$ .

**5.E.117.** The answer here is the square with sides of length  $c$ .

**5.E.118.** We get  $h = \frac{4}{3}R$  and  $r = \frac{2\sqrt{2}}{3}R$ .

**5.E.119.** This triangle is the equilateral triangle, with area  $\sqrt{3}p^2/36$ .

**5.E.120.** The desired points have coordinates  $P = [2, -1/2]$ , and  $Q = [-2, -1/2]$ , respectively. Try to illustrate the situation via Sage.

**5.E.122.** To confirm the answer presented for  $x_0$  in [5.E.121](#) we may instead try to locate the global minimum of the function

$$g(x) = \frac{1}{f(x)} = \frac{x^2 + (b-h)(a-h)}{x(b-a)} = \frac{x}{b-a} + \frac{(b-h)(a-h)}{x(b-a)}, \quad x \in (0, +\infty).$$

This can be done with the help of the AM-GM inequality that we met in Chapter 1, that is,

$$\frac{y_1 + y_2}{2} \geq \sqrt{y_1 y_2}, \quad y_1, y_2 \geq 0,$$

where the equality occurs if and only if  $y_1 = y_2$ . The choice

$$y_1(x) = \frac{x}{b-a}, \quad y_2(x) = \frac{(b-h)(a-h)}{x(b-a)}$$

then gives

$$g(x) = y_1(x) + y_2(x) \geq 2 \sqrt{y_1(x) y_2(x)} = \frac{2}{b-a} \sqrt{(b-h)(a-h)}.$$

Therefore, if there is a number  $x > 0$  for which  $y_1(x) = y_2(x)$ , then the function  $g$  has the global minimum at  $x$ . In particular, the equation

$$y_1(x) = y_2(x), \quad \text{i. e.} \quad \frac{x}{b-a} = \frac{(b-h)(a-h)}{x(b-a)}$$

has a unique positive solution given by  $x_0 = \sqrt{(b-h)(a-h)}$ .

**5.E.126.** The inscribed rectangle has sides of lengths  $x$ ,  $\sqrt{3}/2(a-x)$ , thus its area is  $\sqrt{3}/2(a-x)x$ . The maximum occurs for  $x = a/2$ , hence the greatest possible area is  $(\sqrt{3}/8)a^2$ .

**5.E.127.** The dimensions of the pool are  $4 \text{ m} \times 4 \text{ m} \times 2 \text{ m}$ .

**5.E.128.** The answer is  $28 = 24 + 4$ .

**5.E.129.** The answer is  $a = 1$ .

**5.E.131.** (a) The first claim is obviously true. We do not need any absolute convergence to handle linear combinations. This is a direct consequence of the properties of limits.

(b) This statement is clearly false, in general. Consider the harmonic series  $\sum_{n=1}^{\infty} \frac{1}{n}$ . We saw in 5.D.9 that the series  $\sum_{n=1}^{\infty} \frac{1}{n^2}$  converges, while the harmonic series does not.

(c) Let us divide the series into a sum of two. The first one will collect the terms with  $|a_n| > n|a_n|^2$ , while the other sum the remaining terms. In the first case,  $|a_n| < \frac{1}{n}$ , and thus  $|\frac{a_n}{n}| < \frac{1}{n^2}$ . Consequently, this part of the series converges absolutely. The remaining part must converge, as well, as a consequence of the comparison with the converging series:  $|\frac{a_n}{n}| \leq |a_n|^2$ . As a sum of two absolutely convergent series, the initial series is absolutely convergent, too. Thus the claim is true.

**5.E.135.** (a) The ratio test gives

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n \rightarrow \infty} \frac{2^{(n+1)^2} \cdot n!}{(n+1)! \cdot 2^{n^2}} = \lim_{n \rightarrow \infty} \frac{2^{2n+1}}{n+1} = \lim_{n \rightarrow \infty} \frac{2 \cdot 4^n}{n+1} = +\infty.$$

Thus the series does not converge.

(b) The sequence  $\left(\frac{2}{\sqrt{3n}}\right)_{n \in \mathbb{Z}_+}$  is decreasing, while recall that and the function  $f(x) = \arctan(x)$  is increasing on the whole real axis. So, the sequence  $\left(\arctan\left(\frac{2}{\sqrt{3n}}\right)\right)_{n \in \mathbb{Z}_+}$  is decreasing. Thus, it is an alternating series such that the sequence of the absolute values of its terms is decreasing. According to the Leibniz criterion, such an alternating series converges if and only if the sequence of its terms converges to zero, and this is satisfied:  $\lim_{n \rightarrow \infty} \arctan\left(\frac{2}{\sqrt{3n}}\right) = \arctan(0) = 0$  and hence

$$\lim_{n \rightarrow \infty} \left( (-1)^{n+1} \arctan\left(\frac{2}{\sqrt{3n}}\right) \right) = 0.$$

So the series  $T$  is convergent.

**5.E.137.** The required value is  $3/2$ .

**5.E.138.** The first series sums to 3 and the second sums to  $9/4$ .

**5.E.139.** We get  $\alpha > 0$ ,  $\beta \in \{-2, -1, 0, 1, 2\}$ , and  $\gamma \in (-\infty, -1) \cup (1, +\infty)$ .

**5.E.140.** This series is absolutely convergent.

**5.E.141.** The answer is  $\alpha \in [0, 1)$ .

**5.E.144.** We see that  $A(x) = \frac{1}{2} \ln \frac{1+x}{1-x}$  and  $B(x) = \frac{1+x}{(1-x)^3}$ , respectively.

**5.E.145.** For example:  $a_n = n/3$ ,  $b_n = n/2$ , with  $n \in \mathbb{N}$ .

**5.E.146.** The former series converges absolutely; the latter converges conditionally.

**5.E.147.** Yes, it does.

**5.E.148.** The answer is  $p \in \mathbb{R}$ .

**5.E.149.** This series is convergent.

**5.E.150.** The answer is  $S = 2/9$ .



## Differential and integral calculus

*we already have the menagerie, but what shall we do with it?  
– we'll learn to control it...*



### A. Derivatives of higher orders

*In the previous chapter we briefly explained the use of second order derivatives in the study of local extremes (see 5.C.28 for example). In this chapter we will study higher order derivatives in many details and derive such applications in a more systematic way. Notice that a higher order derivative refers to the repeated process of taking derivatives of derivatives, a procedure also known as “successive differentiation”.*

*As in Chapter 5, we will denote the second derivative of a function  $f$  by  $f'' = (f')'$ , and for derivatives of third or higher order we will write  $f^{(3)} = (f'')'$ ,  $f^{(4)} = (f^{(3)})'$ , etc. The bracket in the notation  $f^{(n)} = (f^{(n-1)})'$  is necessary to distinguish it from the  $n$ th power  $f^n$  of  $f$ . When for a function  $f$  we can consider arbitrarily many continuous derivatives, then the function is said to be smooth, see also the theoretical column for the notion of smooth functions (e.g., 6.1.9).*

*At this point it is also important to recall that in Sage the  $n$ th order derivative of a given function  $f$  can be computed via the command `diff(f, x, n)`. For the examples described below we advise the reader to use Sage and verify the computations, especially for the problems where a Sage implementation is not included.*

In the previous chapter, we were working either with an extremely large class of functions (for example, all continuous, all differentiable), or with only particular functions, (for example exponential, trigonometric, polynomial). However we had very few tools. We indicated how to discuss the local behaviour of functions near a given point by linear approximation. We learned how to measure instantaneous changes by differentiation.

Now we derive several results that will allow us to work with functions more easily when modeling real problems. We also deal with the task of summing infinitely many “infinitely small” changes, in particular, how to “integrate”. In the last part of the chapter we come back to series of functions. We also add useful techniques, how to deal with extra parameters in the functions, and we introduce some further integration concepts briefly.

### 1. Differentiation

**6.1.1. Higher order derivatives.** If the first derivative  $f'(x)$  of a function of one real variable has a derivative  $(f')'(x_0)$  at the point  $x_0$ , we say that the *second derivative* of function  $f$  (or second order derivative) exists. Then we write  $f''(x_0) = (f')'(x_0)$  or  $f^{(2)}(x_0)$ . A function  $f$  is *two times differentiable* on some interval, if it has a second derivative at each of its points. Derivatives of higher orders are defined inductively:



#### $k$ TIMES DIFFERENTIABLE FUNCTIONS

A function  $f$  of one real variable is *differentiable  $k$  times* at the point  $x_0$  for some natural number  $k > 1$ , if it is differentiable  $(k - 1)$  times on some neighbourhood of the point  $x_0$  and its  $(k - 1)$ -st derivative has a derivative at the point  $x_0$ . We write  $f^{(k)}(x)$  for the  $k$ -th derivative of the function  $f(x)$ .

If derivatives of all orders exist on an interval  $A$ , we say the function  $f$  is *smooth* or *infinitely differentiable* on  $A$ .

We use the notation *class of functions*  $C^k(A)$  for functions with continuous  $k$ -th derivative on the interval  $A$ , where  $k$  can attain values  $0, 1, \dots, \infty$ . Often we write only  $C^k$ , if the domain is known from the context. When  $k = 0$ ,  $C^0$  means *continuous functions*.

**6.A.1.** Determine the derivatives given below.

- |   |   |
|---|---|
| (1) $(x^2 \sin(x))''$ , $x \in \mathbb{R}$  | (5) $(x^x)''$ , $x > 0$                             |
| (2) $\left(\frac{2x}{4x+3}\right)''$ , $x \in \mathbb{R} \setminus \{-\frac{3}{4}\}$                  | (6) $(x^n)^{(n)}$ , $x \in \mathbb{R}$              |
| (3) $\left(\ln \sqrt{\frac{x^2+1}{x^2-1}}\right)''$ , with<br>$x \in (-\infty, -1) \cup (1, +\infty)$ | (7) $\left(\frac{x}{\ln(x)}\right)^{(3)}$ , $x > 0$ |
| (4) $(\tan(x))''$ , with<br>$x \in \mathbb{R} \setminus \{\frac{\pi}{2} + n\pi : n \in \mathbb{N}\}$  | (8) $(1/\sqrt{x(x-2)})''$ , with<br>$0 < x < 2$ .   |

**Solution.** (1) This is based on the product rule  $(fg)' = f'g + fg'$ . Applying this rule successively we obtain

$$(x^2 \sin(x))'' = \left((x^2 \sin(x))'\right)' = (2x \sin(x) + x^2 \cos(x))' = 2 \sin(x) + 4x \cos(x) - x^2 \sin(x).$$

(2) In this case we will apply twice the quotient rule  $\left(\frac{f}{g}\right)' = \frac{gf' - fg'}{g^2}$ . We get  $\left(\frac{2x}{4x+3}\right)' = \frac{2(4x+3) - 8x}{(4x+3)^2} = \frac{6}{(4x+3)^2}$ , hence

$$\begin{aligned} \left(\frac{2x}{4x+3}\right)'' &= \left(\frac{6}{(4x+3)^2}\right)' = \frac{6}{(4x+3)^3} \\ &= -\frac{48(4x+3)}{(4x+3)^4} = -\frac{48}{(4x+3)^3}. \end{aligned}$$

(3) For any  $x \in (-\infty, -1) \cup (1, +\infty)$  we can write

$$\begin{aligned} \ln \sqrt{\frac{x^2+1}{x^2-1}} &= \ln \left(\frac{x^2+1}{x^2-1}\right)^{\frac{1}{2}} = \frac{1}{2} \ln \left(\frac{x^2+1}{x^2-1}\right) \\ &= \frac{1}{2} [\ln(x^2+1) - \ln(x^2-1)]. \quad (*) \end{aligned}$$

This relation has many advantages. For instance, one can easily deduce that the given function is differentiable (as the difference of two differentiable functions). On the other hand, recall that a positive differentiable function  $f$  satisfies  $(\ln(f(x)))' = f'(x)/f(x)$  for all  $x$  in its domain. Hence, a combination of this rule with (\*) gives

$$\left(\ln \sqrt{\frac{x^2+1}{x^2-1}}\right)' = \frac{1}{2} \left[\frac{2x}{x^2+1} - \frac{2x}{x^2-1}\right] = -\frac{2x}{x^4-1}.$$

Now we can easily compute also the second derivative:

$$\left(\ln \sqrt{\frac{x^2+1}{x^2-1}}\right)'' = \left(-\frac{2x}{x^4-1}\right)' = \frac{2(1+3x^4)}{(x^4-1)^2}.$$

(4) For this you should apply the quotient rule twice. In particular, recall by 5.E.76 that  $(\tan(x))' = \frac{1}{\cos^2(x)}$ . Therefore,

$$\begin{aligned} \tan''(x) &\equiv (\tan(x))'' \equiv \frac{d^2}{dx^2} \tan(x) = \left(\frac{1}{\cos^2(x)}\right)' \\ &= 2 \frac{\sin(x)}{\cos^3(x)} = 2 \frac{\tan(x)}{\cos^2(x)}. \end{aligned}$$

We illustrate the concept of higher order derivatives with polynomials. Because the derivative of a polynomial is a polynomial with the degree one less than the original one, after a finite number of differentiations we obtain the zero polynomial. If  $k$  is the degree of the polynomial, then exactly  $k + 1$  differentiations yields zero. Then derivatives of all orders exist, hence  $f \in C^\infty(\mathbb{R})$ .

In the spline construction, see 5.1.9, we took care that the resulting functions belong to the class  $C^2(\mathbb{R})$ . Their third derivatives are piece-wise constant functions. That is why the splines do not belong to  $C^3(\mathbb{R})$ , even though all their higher order derivatives are zero in all of the inner points of all single intervals in the interpolation. Think this example through in detail!

The next assertion is a combinatorial corollary of Leibniz's rule for differentiation of a product of two functions:

**Lemma.** If two functions  $f$  and  $g$  have derivatives of order  $k$  at the point  $x_0$ , then their product also has the derivative of order  $k$  and the following equality holds:

$$(f \cdot g)^{(k)}(x_0) = \sum_{i=0}^k \binom{k}{i} f^{(i)}(x_0) g^{(k-i)}(x_0).$$

**PROOF.** For  $k = 0$ , the statement is trivial. For  $k = 1$  it is Leibniz's product rule. Suppose equality holds for some  $k$ . Differentiate the right hand side and use Leibniz's rule to obtain the expression

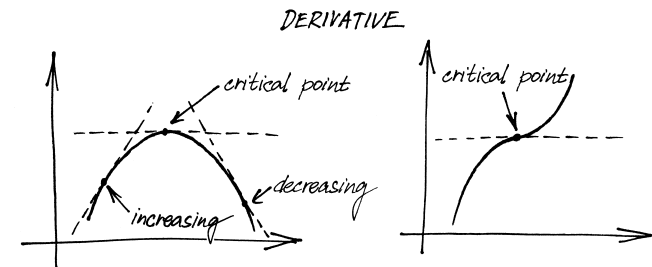
$$\sum_{i=0}^k \binom{k}{i} \left( f^{(i+1)}(x_0) g^{(k-i)}(x_0) + f^{(i)}(x_0) g^{(k-i+1)}(x_0) \right).$$

In this new sum, the sum of orders of the derivatives of products in all summands is  $k + 1$  and the coefficients of  $f^{(j)}(x_0) g^{(k+1-j)}(x_0)$  are the sums of binomial coefficients  $\binom{k}{j-1} + \binom{k}{j} = \binom{k+1}{j}$ .  $\square$

**6.1.2. The meaning of second derivative.** We have already



seen that the first derivative of a function is its linear approximation in the neighbourhood of a given point. The sign of a nonzero derivative determines whether the function is increasing or decreasing at the point  $x_0$ . The points where the first derivative is zero are called the *critical points* or *stationary points* of the given function.



If  $x_0$  is a critical point of function  $f$ , there are several possibilities for the behaviour of the function  $f$  in the neighbourhood of  $x_0$ . Consider the behaviour of the function  $f(x) = x^n$  in the neighbourhood of zero for different

(5) In 5.C.7 (d) we proved that  $(x^x)' = (\ln(x) + 1)x^x$ . Thus, for the second derivative we obtain

$$\begin{aligned} (x^x)'' &= \left( (x^x)' \right)' = \left( (\ln(x) + 1) e^{x \ln(x)} \right)' \\ &= \left( \ln(x) e^{x \ln(x)} \right)' + \left( e^{x \ln(x)} \right)' \\ &= \frac{1}{x} e^{x \ln(x)} + \ln(x) (\ln(x) + 1) e^{x \ln(x)} \\ &\quad + (\ln(x) + 1) e^{x \ln(x)} \\ &= \frac{x^x}{x} + x^x (\ln^2(x) + 2 \ln(x) + 1) \\ &= x^{x-1} + x^x (\ln(x) + 1)^2. \end{aligned}$$

(6) In this case we simply state the result and leave a formal computation for practice:  $\left(\frac{x}{\ln x}\right)^{(3)} = \frac{1}{x^2(\ln x)^2} - \frac{6}{x^2(\ln x)^4}$ .

(7) We see that

$$(x^n)^{(n)} = [(x^n)']^{(n-1)} = (nx^{n-1})^{(n-1)} = \dots = n!.$$

The last step of a formal proof includes an induction over  $n$ , which we leave as an exercise.

(8) For this case we just present the result, and leave the proof for practice:

$$\begin{aligned} \left(1/\sqrt{x(x-2)}\right)'' &= \frac{(2x^2 - 4x + 3)\sqrt{x^2 - 2x}}{x^6 - 6x^5 + 12x^4 - 8x^3} \\ &= \frac{2x^2 - 4x + 3}{\sqrt{(x-2)x(x-2)^2x^2}}. \end{aligned}$$

□

**6.A.2.** Use Sage to verify your answers in 6.A.1.

**Solution.** For most of the cases the answer is based on the command `diff(f, x, m)`, mentioned above. For instance, type

```
show(diff(x^2*sin(x), x, 2).full_simplify())
show(diff((2*x)/(4*x+3), x, 2).full_simplify())
f(x)=ln(sqrt((x^2+1)/(x^2-1)))
show(diff(f(x), x, 2).full_simplify())
show(diff(tan(x), x, 2))
show(diff(x^x, x, 2))
show(diff(x/ln(x), x, 3))
g(x)=1/sqrt(x*(x-2))
show(diff(g(x), x, 2).factor())
```

Notice the function `full_simplify()` (or `factor()`) was used to simplify the given expressions and make direct the verification, whenever necessary.

Finally observe also that the cell

```
[diff(x^n, x, n) for n in range(1, 100)]
```

gives a verification of the relation  $(x^n)^{(n)} = n!$ , presented in (7), for many values of  $n$ . However, a more precise program that offers a more practical demonstration of the theoretical result is as follows:"

```
# Define the variable x
x = var('x')
# Define a specific integer value for n
n = 4 # Change this to test different cases
```

$n$ . For odd  $n > 0$ ,  $f(x)$  will be increasing on all  $\mathbb{R}$ , while for even  $n$  it will be decreasing for  $x < 0$  and increasing for  $x > 0$ . In the latter case, the function will attain its minimal value among points in the (sufficiently small) neighbourhood of  $x_0 = 0$ .

The same argument applies to the function  $f'$ . If the second derivative is nonzero, its sign determines the behaviour of the first derivative. At the critical point  $x_0$  the derivative  $f'(x)$  is increasing if the second derivative is positive and decreasing if the second derivative is negative. If increasing, it is necessarily negative to the left of the critical point and positive to the right of it. In that case,  $f$  is decreasing to the left of the critical point and increasing to the right of it. So  $f$  attains its minimal value among all points from a (sufficiently small) neighbourhood of  $x_0$  at  $x_0$ .

On the other hand, if the second derivative is negative at  $x_0$ , the first derivative is decreasing. Thus the first derivative is negative to the left of  $x_0$  and positive to the right of it.  $f$  then attains its maximal value at  $x_0$  among all values in a neighbourhood of  $x_0$ .

A function which is differentiable on  $(a, b)$  and continuous on  $[a, b]$  has an absolute maximum and minimum on this interval. Both can be attained only at the boundary of the interval or at a point where the derivative is zero. Thus critical points may be sufficient for finding extremes. Second derivatives help to determine the type of extreme, if nonzero.

For a more precise discussion of the latter phenomena we consider higher order polynomial approximations of the functions. We then return to the qualitative study of the behaviour of functions with new tools.

**6.1.3. Taylor expansion.** As a surprisingly easy use of Rolle's theorem we derive an extremely important result. It is called the *Taylor expansion with remainder*<sup>1</sup>.



Consider the power series centered at  $a$ ,

$$S(x) = \sum_{n=0}^{\infty} a_n(x-a)^n.$$

Recall, power series can be differentiated term by term, cf. 5.4.10.

Differentiate the series  $S(x)$  repeatedly, to get the power series

$$S^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)a_n(x-a)^{n-k}.$$

Put  $x = a$ . Then  $S^{(k)}(a) = k! a_k$ . We can read the last statement as an equation for  $a_k$  and rewrite the original series as

$$S(x) = \sum_{n=0}^{\infty} \frac{1}{k!} S^{(k)}(a)(x-a)^n.$$

<sup>1</sup>Brook Taylor was an English mathematician (1685-1731) best known for his formalization of the polynomial approximations of functions, recognized by Lagrange as the "main foundation of differential calculus"

```
# Define the function f(x) = x^n
f = x^n
# Compute the n-th derivative
n_th_derivative = f.diff(x, n)
# Simplify the result
n_th_derivative_s = n_th_derivative.simplify()
# Print the result
print(f"The {n}-th derivative of x^{n} is:",
      n_th_derivative_s)
# Verify that the result matches n!
expected_result = factorial(n)
print(f"Expected result ({n}!):",
      expected_result)
# Check if the computed result matches n!
if n_th_derivative_s == expected_result:
    print("The result matches n!")
else:
    print("The result does not match n!.")
```

□

**6.A.3.** Use the generalized Leibniz rule (see lemma in 6.1.1) to compute the fourth-order derivative of the functions:

- (a)  $h(x) = x \cos(x)$ , (c)  $\ell(x) = x^3 \ln(x)$ ,
- (b)  $k(x) = e^{4x} x^4$ , (d)  $m(x) = e^x \arctan(x)$ .

Next describe a solution in Sage.

**Solution.** The generalized Leibniz’s rule for the 4th-order derivative of a product  $fg$  gives

$$\begin{aligned} (f(x)g(x))^{(4)} &= \binom{4}{0} f^{(0)}(x)g^{(4)}(x) + \binom{4}{1} f'(x)g^{(3)}(x) \\ &+ \binom{4}{2} f''(x)g''(x) + \binom{4}{3} f^{(3)}(x)g'(x) + \binom{4}{4} f^{(4)}(x)g(x) \\ &= f(x)g^{(4)}(x) + 4f'(x)g^{(3)}(x) + 6f''(x)g''(x) \\ &+ 4f^{(3)}(x)g'(x) + f^{(4)}(x)g(x). \quad (\#) \end{aligned}$$

Now, for case (a) we have  $h(x) = f(x)g(x)$  with  $f(x) = x$  and  $g(x) = \cos(x)$ , respectively. Moreover,  $g'(x) = -\sin(x)$ ,  $g''(x) = -\cos(x)$ ,  $g'''(x) = \sin(x)$ ,  $g^{(4)}(x) = \cos(x)$ ,  $f'(x) = 1$ , and  $f^{(k)}(x) = 0$  for all  $k \geq 2$ . Thus, an application of (#) gives  $h^{(4)}(x) = x \cos(x) + 4 \sin(x)$ .

A solution in Sage is based on the cell

```
h(x)=x*cos(x); show(diff(h,x,4))
```

Or, you may use the alternative

```
((diff(h(x),x,i) for i in [1..4]))
```

which gives us all the derivatives  $h'(x)$ ,  $h''(x)$ ,  $h^{(3)}(x)$  and  $h^{(4)}(x)$ . Similarly are treated the rest cases. □

**6.A.4. Generalized Leibniz rule and Sage.** For two symbolic functions  $f, g$  in Sage write a short program to implement the identity (#) presented in 6.A.3. Then verify the results for the functions  $h, k, \ell, m$ , by using this program, and hence independently of the command `diff(f * g, x, 4)`.

**Solution.** Recall from 5.C.8 that in Sage we can easily introduce symbolic functions by typing

Suppose  $f$  is a smooth function instead of a power series. We search for a good approximation by polynomials in the neighbourhood of a given point  $a$ .

TAYLOR POLYNOMIAL OF FUNCTION  $f$

For a  $k$  times differentiable (real or complex valued) function  $f$  of one real variable, define its *Taylor polynomial of  $k$ -th degree* centered at  $a$  by the formula

$$\begin{aligned} T_a^k f(x) &= f(a) + f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2 \\ &+ \frac{1}{6} f^{(3)}(a)(x - a)^3 + \dots + \frac{1}{k!} f^{(k)}(a)(x - a)^k. \end{aligned}$$

The mean value theorem is used to show how good this approximation of  $f$  is.

TAYLOR EXPANSION WITH A REMAINDER

**Theorem.** Let  $f(x)$  be a function that is  $k$  times differentiable on the interval  $(a, b)$  and continuous on  $[a, b]$ . Then for all  $x \in (a, b)$  there exists a number  $c \in (a, x)$  such that

$$\begin{aligned} f(x) &= T_a^{k-1} f(x) + \frac{1}{k!} f^{(k)}(c)(x - a)^k \\ &= f(a) + f'(a)(x - a) + \dots \\ &+ \frac{1}{(k-1)!} f^{(k-1)}(a)(x - a)^{k-1} + \frac{1}{k!} f^{(k)}(c)(x - a)^k. \end{aligned}$$

**PROOF.** For fixed  $x$  define the remainder  $R$ , by



$$f(x) = T_a^{k-1} f(x) + R.$$

Then  $R = \frac{1}{k!} r(x - a)^k$  for a suitable number  $r$  (dependent on  $x$ ). Consider the function  $F(\xi)$  defined by

$$F(\xi) = \sum_{j=0}^{k-1} \frac{1}{j!} f^{(j)}(\xi)(x - \xi)^j + \frac{1}{k!} r(x - \xi)^k.$$

By the Leibniz rule, its derivative (here  $x$  is considered as a constant parameter) is

$$\begin{aligned} F'(\xi) &= f'(\xi) + \sum_{j=1}^{k-1} \left( \frac{1}{j!} f^{(j+1)}(\xi)(x - \xi)^j - \frac{1}{(j-1)!} f^{(j)}(\xi)(x - \xi)^{j-1} \right) \\ &\quad - \frac{1}{(k-1)!} r(x - \xi)^{k-1} \\ &= \frac{1}{(k-1)!} f^{(k)}(\xi)(x - \xi)^{k-1} - \frac{1}{(k-1)!} r(x - \xi)^{k-1} \\ &= \frac{1}{(k-1)!} (x - \xi)^{k-1} (f^{(k)}(\xi) - r), \end{aligned}$$

because the expressions in the sum cancel each other out sequentially. Now it suffices to notice that  $F(a) = F(x) = f(x)$  (recall that  $x$  is an arbitrarily chosen but fixed number from the interval  $(a, b)$ ). According to Rolle’s theorem there exists a number  $c$ ,  $a < c < x$  such that  $F'(c) = 0$ . That is the desired relation. □

```
var("x"); f=function("f")(x)
```

This, in combination with the `substitute_function`, gives rise to an alternative approach to obtain the first derivative of a given function. For instance, for  $f(x) = e^x$  we just need to add the code

```
f1=f.diff(x)
show(f1.substitute_function(f==exp(x)))
```

Check yourself that this block returns the first derivative of  $f(x) = e^x$ .

We are now ready to verify (‡), which can be done for example by the cell

```
f=function("f")(x)
g=function("g")(x)
f1=diff(f, x); f2=diff(f, x, 2)
f3=diff(f, x, 3); f4=diff(f, x, 4)
g1=diff(g, x); g2=diff(g, x, 2)
g3=diff(g, x, 3); g4=diff(g, x, 4)
h4x=f*g4+4*f1*g3+6*f2*g2+4*f3*g1+f4*g
show(h4x)
```

This block includes all the derivatives that one needs to encode (‡) inside the program, which we did here by the syntax named “h4x”. To confirm the expression of  $h^{(4)}(x)$  given in 6.A.3, it now suffices to add the code


```
show(h4x.substitute_function(f==x, g==cos(x)))
```

The rest cases can be confirmed similarly and left for practice. □

**6.A.5.** Let  $n \in \mathbb{N}$  be arbitrary. Find the  $n$ th derivative of function  $f(x) = \ln\left(\frac{1+x}{1-x}\right)$ , where  $x \in (-1, 1)$ . ○

**6.A.6.** Compute the 12th derivative of the function  $f(x) = e^{2x} + \cos(x) + x^{10} - 5x^7 + 6x^3 - 7x + 3$ , with  $x \in \mathbb{R}$ . ○

**6.A.7.** Compute the 26th derivative of the function  $f(x) = \sin(x) + x^{23} - x^{18} + 5x^{11} - 3x^8 + e^{2x}$ , with  $x \in \mathbb{R}$ . ○

 *Taylor polynomials extend the idea of linearization that we briefly discussed in Chapter 5. They provide high-order approximations of a function  $f$  by polynomials in the neighbourhood of a point (and thus locally). This is because such expansions are based on higher order derivatives of  $f$  at that point. We remark that as the terms of such polynomials increase (and hence their order), the induced approximations are improved.*

*Taylor expansions have many natural applications in mathematics, but also in other sciences (e.g., in physics). In the sequel, first are described exercises related to the computation of Taylor polynomials, and more interesting applications will be discussed hereafter. For our description we adopt the notation from the theoretical section 6.1.3, hence for example we will denote by  $T_a^k f(x)$  the  $k$ th-order Taylor expansion of  $f(x)$  around a point  $a$ .*

A special case of the last theorem is the mean value theorem, as an approximation by the Taylor polynomial of degree zero. See 5.3.9(1).

**6.1.4. Estimates for Taylor expansions.** A simple case of a Taylor expansion is when  $f$  is a polynomial:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0, \quad a_n \neq 0.$$

Because the  $(n + 1)$ -th derivative  $f$  is identically zero, the Taylor polynomial of degree  $n$  has zero remainder, therefore for each  $x_0 \in \mathbb{R}$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \dots + \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n.$$

We can compute all the derivatives easily (for example the last term is always of the form  $a_n(x - x_0)^n$ ).

This result is a very special case of error estimation in Taylor expansion with the remainder. We know in advance that the remainder can be estimated by the size of the derivative, and for polynomials this is identically zero for some order onwards.

More generally, the estimation of the size of the  $k$ -th derivative on some interval can be used to estimate the error on the same interval.

Good examples of an expansion of an arbitrary degree are provided by the trigonometric functions  $\sin$  and  $\cos$ . By iterating the differentiation of the function  $\sin x$  we always have either sine or cosine with some signs. The absolute values do not exceed one. Thus we obtain a direct estimation of the speed of convergence of the power series

$$|\sin x - (T_0^k \sin)(x)| \leq \frac{|x|^{k+1}}{(k+1)!}.$$

This shows that for  $x$  much smaller than  $k$  the error is small, but for  $x$  comparable with  $k$  or bigger it may be large. In the figure, compare the approximation of the function  $\cos x$  by a Taylor polynomial of degree 68 in paragraph 5.4.12 on the page 429.

As mentioned in the introduction of the discussion of Taylor expansion of functions, if we start with a power series  $f(x)$  centered at  $a$ , then its partial sums coincide with Taylor polynomials  $T_a^k f(x)$ . The next statement is one of the simple formulations of the converse implication. This is when the given function  $f(x)$  is actually a power series on some neighbourhood of the given point  $a$ .

#### TAYLOR'S THEOREM

**Theorem.** Assume that the function  $f(x)$  is smooth on the interval  $(a - b, a + b)$  and all of its derivatives are bounded uniformly by a constant  $M > 0$ . So

$$|f^{(k)}(x)| \leq M, \quad k = 0, 1, \dots, x \in (a - b, a + b).$$

Then the power series  $S(x) = \sum_{n=0}^{\infty} \frac{1}{k!} f^{(k)}(a)(x - a)^n$  converges on the interval  $(a - b, a + b)$  to  $f(x)$ .

**6.A.8.** For the elementary functions  $\sin(x)$ ,  $\cos(x)$ ,  $e^x$  and  $\ln(x + 1)$  compute the third order Taylor expansion based on  $a = 0$ . ○

**6.A.9.** Determine the Taylor expansion  $T_1^3 f(x)$  for the functions  $f(x) = \frac{e^x}{x}$  and  $f(x) = \arctan(x)$ , with  $x \in \mathbb{R} \setminus \{0\}$  and  $x \in \mathbb{R}$ , respectively. Then verify your answers in Sage and moreover check the accuracy of the approximation near the point  $a$ .

**Solution.** The third-order Taylor expansion around the point  $a = 1$  is given by

$$T_1^3 f(x) = f(1) + f'(1)(x-1) + \frac{f''(1)}{2}(x-1)^2 + \frac{f'''(1)}{6}(x-1)^3.$$

For the first case we have  $f(1) = e$ , and

$$\begin{aligned} f'(1) &= \left. \frac{e^x}{x} - \frac{e^x}{x^2} \right|_{x=1} = 0, \\ f^{(2)}(1) &= \left. \frac{e^x}{x} - 2\frac{e^x}{x^2} + 2\frac{e^x}{x^3} \right|_{x=1} = e, \\ f^{(3)}(1) &= \left. \frac{e^x}{x} - 3\frac{e^x}{x^2} + 6\frac{e^x}{x^3} - 6\frac{e^x}{x^4} \right|_{x=1} = -2e. \end{aligned}$$

Thus it follows that  $T_1^3 \frac{e^x}{x} = e + \frac{e}{2}(x-1)^2 - \frac{e}{3}(x-1)^3$ .

In Sage in order to compute the Taylor expansion  $T_a^k f(x)$  of a given function  $f(x)$  around a point  $a$ , you can either apply the command `taylor(f(x), x, a, k)`, or type `f.taylor(x, a, k)`. Thus, a verification of the derivatives and the Taylor expansion presented above is obtained by the cell

```
f(x)=e^x/x
a=diff(f, x, 2); show(a); show(a(1))
b=diff(f, x, 3); show(b); show(b(1))
show(taylor(f(x), x, 1, 3))
```

In order to check the accuracy of the approximation near the point  $a = 1$ , we can check the values of  $f$  and  $T_1^3 f$  at a point close to  $a = 1$ , say at  $x = 1.1$ . The block

```
f(x)=exp(x)/x; T(x)=taylor(f(x), x, 1, 3)
show("Value of the function at x=1.1:",
     round(f(x=1.1), 10))
show("3rd degree Taylor pol. at x=1.1:",
     round(T(x=1.1), 10))
```

returns the desired answer, namely:

- Original value of the function at  $x = 1.1$ : 2.7310600218,
- 3rd degree Taylor polynomial at  $x = 1.1$ : 2.7309671437.

The function  $\arctan(x)$  can be treated in the same way. For instance, the syntax `show(taylor(arctan(x), x, 1, 3))` gives us the following expression.

$$T_1^3 \arctan(x) = \frac{\pi}{4} + \frac{1}{2}(x-1) - \frac{1}{4}(x-1)^2 + \frac{1}{12}(x-1)^3.$$

**PROOF.** The proof is identical with the special case of function  $\sin x$  above. Except that the universal bound by 1 is replaced by  $M$ , and thus the estimate of the remainders are

$$|f(x) - (T_a^k f)(x)| \leq \frac{M}{(k+1)!} |x|^{k+1}. \quad \square$$

**6.1.5. Local behaviour of functions.** With the Taylor expansions at hand, let us return to the local or global behaviour of real functions of one real variable. We have seen that the sign of the first derivative of a differentiable function determines whether it is increasing or decreasing on some neighbourhood of the given point. If the derivative is zero, it is one of the so called critical points, but we do not know about the function much without looking at higher derivatives.



We encountered the importance of the second derivative when analysing the critical points, see 6.1.2. Now we generalize the discussion of critical points for all orders. First we deal with the local extremes of functions.

In the following we consider real functions with a sufficiently high number of continuous derivatives, without specifically stating this assumption.

The point  $a$  in domain of  $f$  is a *critical point of order  $k$*  if and only if

$$f'(a) = \dots = f^{(k)}(a) = 0, \quad f^{(k+1)}(a) \neq 0.$$

Suppose  $f^{(k+1)}(a) > 0$  and  $f \in C^{k+1}$ . Then this continuous derivative is positive on a certain neighbourhood  $\mathcal{O}(a)$  of the point  $a$  as well. In that case, the Taylor expansion with the remainder gives

$$f(x) = f(a) + \frac{1}{(k+1)!} f^{(k+1)}(a)(x-a)^{k+1}$$

for all  $x$  in  $\mathcal{O}(a)$ . Because of that, the change of values of  $f(x)$  in a neighbourhood of  $a$  is given by the behaviour of  $(x-a)^{k+1}$ . Moreover, if  $k+1$  is an even number, then the values of  $f(x)$  in such a neighbourhood are necessarily larger than the value  $f(a)$ . So  $a$  is a local minimum. But if  $k$  is even then the values on the left are smaller, while those on right are larger than  $f(a)$ . So an extreme does not occur even locally. On the other hand, the graph of the function  $f(x)$  intersects its tangent  $y = f(a)$  at the point  $[a, f(a)]$  in the latter case, as discussed in more detail below.

Similarly, if  $f^{(k+1)}(a) < 0$ , then it is a local maximum for odd  $k$ , and there is no extreme for even  $k$ .

**6.1.6. Convex and concave functions.** The differentiable function  $f$  is *concave* at  $a$ , if its graph lies completely below the tangent at the point  $[a, f(a)]$  in a neighbourhood of  $a$ . That is,



$$f(x) \leq f(a) + f'(a)(x-a).$$

Similarly  $f$  is *convex* at  $a$ , if its graph is above the tangent at the point  $a$ . That is,

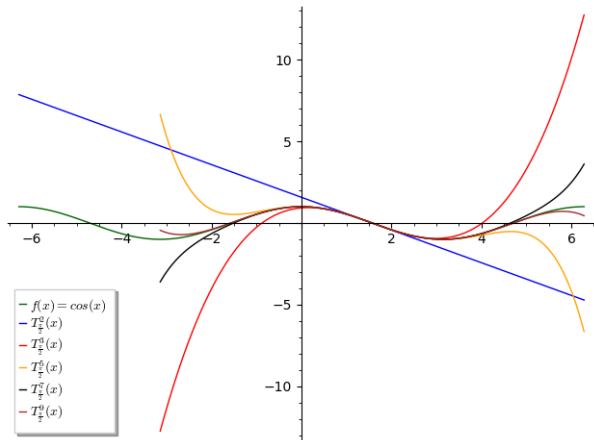
$$f(x) \geq f(a) + f'(a)(x-a).$$

To check the accuracy, one can apply the same method as for  $f(x) = \frac{e^x}{x}$ .  $\square$

**6.A.10.** Determine the Taylor expansion  $T_1^4 f(x)$  where  $f(x) = \ln(x^2)$ , with  $x \in (0, 2)$ . Next use Sage to plot together  $f(x)$  and  $T_1^4 f(x)$ , for  $x$  in the domain of  $f$ .  $\circ$

**6.A.11.** Consider the function  $f(x) = \cos(x)$  with  $x \in \mathbb{R}$ . Use Sage to sketch in one diagram the graph of  $f$  together with the graphs of Taylor polynomials  $T_a^k f(x)$  for  $k = 2, 3, 5, 7, 9$  and  $a = \pi/2$ , for certain intervals of  $\mathbb{R}$ .

**Solution.** Sage allows us to use different colours to present the required graphs, and this makes easier their recognition. Based on this remark, one can present the following figure.



To produce this figure we use the `taylor` function to introduce the required Taylor polynomials. We then sketch their graphs via the `plot` command, which we use with some more advanced options (for example, any polynomial comes with a label, that we did via the `legend_label` option). Let us present the full syntax in a block:

```
f(x)=cos(x)
p=plot(f(x), x, -2*pi, 2*pi,color="green",
legend_label=r"$f(x)=\cos(x)$")
T2(x)=taylor(f(x), x, pi/2, 2)
T3(x)=taylor(f(x), x, pi/2, 3)
T5(x)=taylor(f(x), x, pi/2, 5)
T7(x)=taylor(f(x), x, pi/2, 7)
T9(x)=taylor(f(x), x, pi/2, 9)
p2=plot(T2(x), x, -2*pi, 2*pi,
legend_label=r"$T^2_{\frac{\pi}{2}}(x)$")
p3=plot(T3(x), x, -pi, 2*pi, color="red",
legend_label=r"$T^3_{\frac{\pi}{2}}(x)$")
p5=plot(T5(x), x, -pi, 2*pi, color="orange",
legend_label=r"$T^5_{\frac{\pi}{2}}(x)$")
p7=plot(T7(x), -pi, 2*pi, color="black",
legend_label=r"$T^7_{\frac{\pi}{2}}(x)$")
p9=plot(T9(x), -pi, 2*pi, color="brown",
legend_label=r"$T^9_{\frac{\pi}{2}}(x)$")
show(p+p2+p3+p5+p7+p9)
```

For further practice with Taylor polynomials, check [6.B.40](#) and see also in [Section D](#) for a series of additional tasks (e.g., [6.D.5](#)).  $\square$

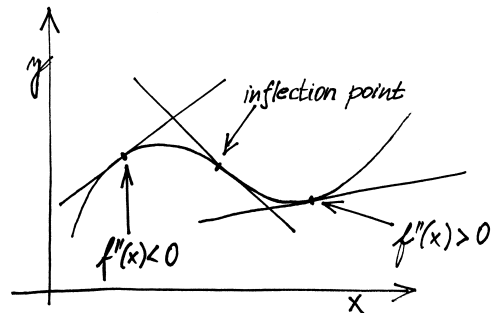
A function is convex or concave on an interval, if it has this property at all its points.

Suppose  $f$  has continuous second derivatives in a neighbourhood of  $a$ . The Taylor expansion of second order with the remainder implies

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2}f''(c)(x - a)^2.$$

Then the function is convex, whenever  $f''(a) > 0$ , and concave whenever  $f''(a) < 0$ .

If the second derivative is zero, we can use derivatives of higher orders. But we can only make the same conclusion if the first other nonzero derivative after the first derivative is of even order. If the first nonzero derivative is of odd order, the points of the graph of the function on opposite sides of some small neighbourhood of the studied point will lie on opposite sides of the tangent at this point.



**6.1.7. Inflection points.** A point  $a$  is called an *inflection point* of a differentiable function  $f$ , if the graph of  $f$  crosses from one side of the tangent in the point  $a$  to the other. The latter discussion on concave and convex functions shows that the inflections can appear only at points with vanishing second derivative.

Suppose  $f$  has continuous third derivatives and write the Taylor expansion of third order with the remainder:

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2}f''(a)(x - a)^2 + \frac{1}{6}f'''(c)(x - a)^3.$$

If  $a$  is a zero point of the second derivative such that  $f'''(a) \neq 0$ , then the third derivative is nonzero on some neighbourhood. Then  $a$  is an inflection point since the second derivative changes the sign at  $a$  and thus the tangent crosses the graph. In that case, the sign of the third derivative determines whether the graph of the function crosses the tangent from the top to the bottom or vice versa.

Moreover, if  $a$  is an isolated zero point of the second derivative and simultaneously an inflection point, then on some small neighbourhood of  $a$  the function is concave on one side and convex on the other. Thus the inflection points are points of the change between concave and convex behaviour of the graph of the function.



So far we have learned how to use Taylor polynomials to approximate smooth functions. Whenever we use an approximation technique, it is important to have a sense for how accurate our approximations are, as for example the Lagrange interpolation error presented in 5.E.9. For Taylor expansions the situation is encoded by the so called Remainder Theorem, which is described in theoretical section 6.1.3. Let us illustrate this by examples.

**6.A.12.** Determine the Taylor polynomial  $T_0^6 f(x)$ , where  $f(x) = \sin(x)$ , and next estimate the error of the approximation at the point  $x = \pi/4$ , according to the Theorem 6.1.3.

**Solution.** It is easy to prove that

$$T_0^6 \sin(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5.$$

According to the statement in 6.1.3, the estimate of the remainder  $R$  at  $a = 0$  is thus given by (the absolute value of) the expression  $\frac{f^{(k+1)}(c)}{(k+1)!}x^{k+1}$ , for some  $c \in (0, \frac{\pi}{4})$  and  $k = 6$ . We compute  $f^{(7)}(x) = -\cos(x)$ , for all  $x \in \mathbb{R}$ , and so at  $x = \pi/4$  we obtain

$$R(\pi/4) = \left| \frac{-\cos(c)\pi^7}{7!4^7} \right| < \frac{1}{7!} \approx 0,0002.$$

□

**6.A.13.** Show that  $e^x \geq 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$ . ○

**6.A.14.** Find the estimation of the error of the approximation

$$\ln(1+x) \approx x - \frac{x^2}{2}$$

for  $x \in (-1, 0)$ . ○

**6.A.15.** Consider the sine function  $f(x) = \sin(x)$ , with  $x \in \mathbb{R}$ . Compute the Taylor polynomial  $T_0^4 f(x)$  and use Sage to plot together the graphs of  $T_0^4 f(x)$  and of  $f(x)$ . Finally, based on your answer estimate  $\sin(1^\circ)$ . ○

**6.A.16.** Compute an approximation of  $\cos \frac{\pi}{10}$  with an error less than  $10^{-5}$ . ○



Taylor series are power series having as partial sums the Taylor polynomials of a given function. Taylor series need not be convergent, and even if the Taylor series of a function  $f$  does converge, its limit may not be equal to  $f(x)$ .

Functions  $f$  whose Taylor series  $\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$  converges to  $f(x)$  for all  $x$  in some open neighbourhood of  $a$ , are said to be “analytic at  $a$ ”. When  $f$  is analytic at any point in an interval  $I \subset \mathbb{R}$ , then  $f$  is said to be “analytic on  $I$ ”, see also in 6.1.9. As we will explain below, for analytic functions the power series representation allows us to solve problems as if the function were a polynomial.

In general not all the functions are analytic, and in particular functions with discontinuities cannot be even expressed as Taylor series. However, most of the functions that



**6.1.8. Asymptotes of graphs of functions.** We introduce one more useful utility for understanding or sketching the graph of a function. We consider the *asymptotes*. These are lines in  $\mathbb{R}^2$  whose distance from the graph of  $f(x)$  converges to zero for  $x \rightarrow x_0$ .

Thus, an asymptote at the improper point  $\infty$  is a line  $y = ax + b$ , which satisfies

$$\lim_{x \rightarrow \infty} (f(x) - ax - b) = 0.$$

An *asymptote with a slope*. If such an asymptote exists, it satisfies

$$\lim_{x \rightarrow \infty} (f(x) - ax) = b$$

Consequently the limit

$$\lim_{x \rightarrow \infty} \frac{f(x)}{x} = a$$

also exists.

Conversely, if the last two limits exist, the limit from the definition of the asymptote exists as well, thus these are sufficient conditions, too.

The asymptote at the improper point  $-\infty$  is defined similarly.

In this way we find all the lines satisfying the properties of asymptotes with slope. It remains to consider lines perpendicular to the  $x$  axis:

The asymptotes at points  $a \in \mathbb{R}$  are lines  $x = a$  such that the function  $f$  has at least one of the one-sided limits at  $a$  has infinite value. They are called *asymptotes without slope*.

The real rational functions have asymptotes at all zero points of the denominator which are not zero points of the numerator as well.

We consider a simple illustrative example: Let

$$f(x) = x + \frac{1}{x}.$$

$f$  has two asymptotes  $y = x$  and  $x = 0$ . Indeed, the one-sided limits from the right and left at zero are clearly  $\pm\infty$ , while the limit  $f(x)/x = 1 + 1/x^2$  is of course 1 at the improper points. Finally the limit of  $f(x) - x = 1/x$  is zero at the improper points.

By differentiating,

$$f'(x) = 1 - x^{-2}, \quad f''(x) = 2x^{-3}.$$

The function  $f'(x)$  has two zero points  $\pm 1$ . At  $x = 1$ ,  $f$  has a local minimum. At  $x = -1$ ,  $f$  has a local maximum. The second derivative has no zero points in all its domain  $(-\infty, 0) \cup (0, \infty)$ , so  $f$  has no inflection points.



arise in applications are usually analytic. In the end of Chapter 5 we met for example the analytic functions

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1},$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k}.$$

In fact it is not hard to prove that the Taylor series of  $\cos(x)$  occurs from the Taylor series of  $\sin(x)$  around the origin, after differentiation.

Another example is induced by the equality

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k.$$

This will be proved later in Section C, where you will have the chance to learn more details on the convergence of Taylor series (see 6.D.47).

Finally, a remarkable common property of all the examples mentioned above, is that the Taylor series in the r.h.s. are centered at  $a = 0$ . Such Taylor series are referred to as “Maclaurin series”. As we will see, from the Maclaurin series we can build many other examples through substitution and series multiplication.

**6.A.17.** Consider the functions

$$f(x) = \ln\left(\frac{1+x}{1-x}\right), \quad g(x) = e^{x^2} + x^2 e^{-2x}$$

with  $x \in (-1, 1)$  and  $x \in \mathbb{R}$ , respectively. Expand them into a Taylor series centered at the origin, that is, into a Maclaurin series, and then present a verification in Sage.

**Solution.** We first analyze the case for  $f(x)$ . We know that  $\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n$ , which implies that

$$\ln(1-x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (-x)^n = -\sum_{n=1}^{\infty} \frac{1}{n} x^n.$$

Thus, for all  $x \in (-1, 1)$  we have the relation

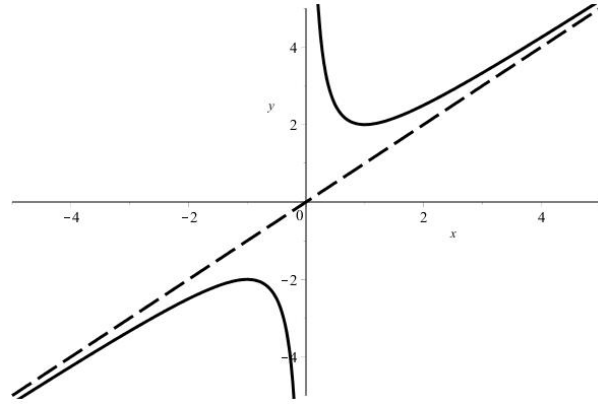
$$\ln\left(\frac{1+x}{1-x}\right) = \ln(1+x) - \ln(1-x) = 2 \sum_{n=1}^{\infty} \frac{x^{2n-1}}{2n-1}.$$

In order to verify this computation in Sage we could type

```
var("n"); assume(x>0)
sum((2/(2*n-1))*x^(2*n-1), n, 1, oo)
```

and this prints out the desired result, i.e.,  $\ln(-(x+1)/(x-1))$ . On the other side, we could use Sage to derive the Taylor series directly, as follows:

```
f(x)=ln((1+x)/(1-x))
tf=taylor(f, x, 0, 10)
T=tf.power_series(QQ)
show(T)
```



**6.1.9. Analytic and smooth functions.** If  $f$  is smooth at  $a$ , the formal power series can be written

$$S(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(a)(x-a)^n.$$

If this power series has a nonzero radius of convergence and simultaneously  $S(x) = f(x)$  on the respective interval, we say that  $f$  is an analytic function at  $a$ . A function is analytic on an interval, if it is analytic at its every point.

One of the most important functions in Mathematics and Physics is the following

GAUSSIAN FUNCTION

The analytic function

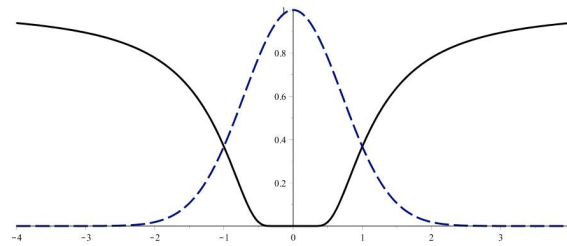
$$g(x) = e^{-x^2}.$$

is called the Gaussian function.

Indeed, the Gaussian is analytic – we simply replace  $x$  with  $-x^2$  in the power series for the exponential  $e^x$  and it is an analytic function on the entire  $\mathbb{R}$ . Its derivatives are easily computed, too. In particular  $g'(0) = 0$  is the only singular point and  $g''(0) = -2$ . Since the limits at  $\pm\infty$  are zero, the number  $g(0) = 1$  is the global maximum of the Gaussian function (see ?? for more detailed comments).

We shall see why the function is so important in Chapter 10 when dealing with probability and statistics.

The Gaussian  $g(x)$  and the function  $f(x) = g(1/x)$  are depicted on the figure,  $f(x)$  is the solid line, while the Gaussian function is dashed.



Notice how firmly the function  $f$  touches the  $x$  axis. We are going to explain this.

This returns the Taylor series corresponding to  $f(x)$ , in the following form

$$2x + \frac{2}{3}x^3 + \frac{2}{5}x^5 + \frac{2}{7}x^7 + \frac{2}{9}x^9 + O(x^{10}).$$

Here, Sage uses the “big O” notation to encode terms of higher order (in our case of order  $\geq 10$ , since we “asked” the Taylor expansion to be of order 10).

For the second function by the identity  $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$  we obtain

$$e^{x^2} = \sum_{n=0}^{\infty} \frac{x^{2n}}{n!}, \text{ for all } x \in \mathbb{R}. \text{ Moreover, } x^2 e^{-2x} =$$

$$x^2 \sum_{n=0}^{\infty} \frac{1}{n!} (-2x)^n = \sum_{n=0}^{\infty} \frac{(-2)^n}{n!} x^{n+2}, \text{ and hence}$$

$$e^{x^2} + x^2 e^{-2x} = \sum_{n=0}^{\infty} \frac{x^{2n} + (-2)^n x^{n+2}}{n!}.$$

The verification in Sage for this case is left for practice.  $\square$

**6.A.18.** Determine the Taylor series centered at the origin for the functions  $f(x) = \frac{1}{(1+x)^2}$ , with  $x \in (-1, 1)$ . Next confirm your result by Sage.  $\circ$

*Maclaurin series are also useful when compute limits, and here is a task for you to perform yourself, see also 6.D.6.*

**6.A.19.** Use the appropriate Maclaurin series to compute the limits  $\lim_{x \rightarrow 0} \frac{x \sin(x) - x^2}{x^4}$ , and  $\lim_{x \rightarrow 0} x^2 \left( e^{-\frac{1}{x^2}} - 1 \right)$ .  $\circ$

Let us now stress the use of derivatives when studying the “local behaviour” of functions, and hence revise the discussion initiated in Chapter 5 (cf. 5.C.15, 5.C.17). This is based on many new notions introduced in ??, and essentially includes the description of the following features of  $f$ :



- the domain and the range;
- parity and periodicity;
- discontinuities and their kind;
- points of intersections with the axes;
- the limits  $\lim_{x \rightarrow \pm\infty} f(x)$ ;
- the first and the second derivatives;
- the critical points (also called stationary points);
- the intervals of monotonicity;
- local and absolute extremes;
- the intervals where  $f$  is convex/concave;
- the points of inflection;
- the horizontal and inclined asymptotes;
- the graph.

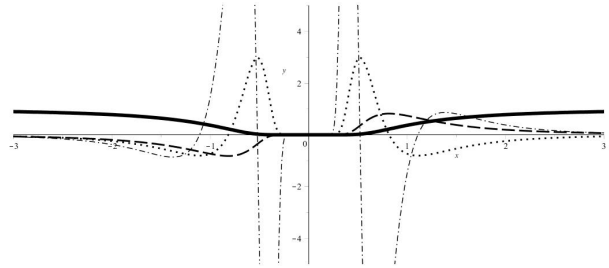
In order to become familiar with these notions, we present many illustrative examples.

Not all smooth functions are analytic. It can be proven that for every sequence of numbers  $a_n$  there is a smooth function, whose derivatives of order  $k$  at a given point  $x_0$  are these numbers  $a_k$ .<sup>2</sup> Now, look more closely at our function



$$f(x) = e^{-1/x^2}.$$

It is a well defined smooth function at all points  $x \neq 0$ . Its limit at  $x = 0$  exists, and  $\lim_{x \rightarrow 0} f(x) = 0$ . By defining  $f(0) = 0$ ,  $f$  is a continuous function for all real  $x$ .



By a direct computation based on L’Hospital’s rule we compute the derivatives of  $f$  (the first three ones are at the picture, guess which line is which!). It suffices to consider only the right derivative, since the function is even.

$$\begin{aligned} f'(0) &= \lim_{x \rightarrow 0^+} \frac{e^{-1/x^2} - 0}{x} = \lim_{x \rightarrow 0^+} \frac{x^{-1}}{e^{1/x^2}} \\ &= \frac{1}{2} \lim_{x \rightarrow 0^+} \frac{x}{e^{1/x^2}} = 0. \end{aligned}$$

By differentiating  $f(x)$  at an arbitrary point  $x \neq 0$ ,  $f'(x) = e^{-1/x^2} \cdot 2x^{-3}$ . By repeated differentiation of the results, there is always a sum of finitely many terms of the form

$$C \cdot e^{-1/x^2} \cdot x^{-j},$$

where  $C$  is an integer and  $j$  is a natural number.

Next, assume it is already proven that the derivative of order  $k$  of  $f(x)$  exists and vanishes at zero. Compute the limit of the expression  $f^{(k)}(x)/x$  for  $x \rightarrow 0^+$ . This is a finite sum of limits of the expressions  $Cx^{-j} e^{-1/x^2} = x^{-j} / e^{1/x^2}$ . All these expressions are of type  $\infty/\infty$ , so L’Hôpital’s rule can be used repeatedly on them. After several differentiations of both the numerator and denominator (and a similar adjustment as above) there remains the same expression in the denominator, while in the numerator the power is non-negative. Thus the expression necessarily has a zero limit at zero, just as in the case of the first derivative above. The same holds for a finite sum of such expressions. So each derivative  $f^{(k)}(x)$  at zero exists with value zero.

<sup>2</sup>This is a special case of the Whitney extension theorem, which says that there is a smooth function on a Euclidean space with prescribed derivatives in all points of a closed set  $A$  if and only if the Taylor theorem estimates are true for the prescription. In the case of one single point  $A$ , the condition is empty. This is relevant for the Taylor theorem for functions of more than one real variable, as in Chapter 8. Hasler Whitney (1907-1989) was a very influential American mathematician.

- 6.A.20.** (a) Find the critical points and the local extremes of the function  $f(x) = \sqrt[3]{x^2} = x^{\frac{2}{3}}$ , with  $x \in \mathbb{R}$ .  
 (b) Provide an example of a function  $g$  satisfying  $g'(0) = 0$  but  $g$  does not attain a local extreme at  $x_0 = 0$ .  
 (c) Provide an example of a function having no local minima or maxima.

**Solution.** (a) The given function  $f$  is continuous in  $\mathbb{R}$  and for any  $x > 0$  we have (see also 5.C.24)

$$f'(x) = (x^{\frac{2}{3}})' = \frac{2}{3}x^{\frac{2}{3}-1} = \frac{2}{3}x^{-\frac{1}{3}} = \frac{2}{3\sqrt[3]{x}}.$$

Comparing the left and right derivatives, one can check that at  $x = 0$  the function  $f$  is *not* differentiable. This can be done quickly via Sage, by the syntax

```
f(x)=x^(2/3)
show(lim((f(x)-f(0))/x, x=0, dir="right"))
show(lim((f(x)-f(0))/x, x=0, dir="left"))
```

which prints out  $+\infty$ , and  $-\infty$ , respectively. Moreover, we see that  $f'(x) = 0$  has no solutions, hence  $f$  has no critical points (in terms of 6.1.5). We already know from Chapter 5 how to obtain a confirmation of the critical points via Sage: just add the cell

```
show(solve(diff(f, x).factor()==0, x))
```

On the other hand, we see that  $f'(x) > 0$  for  $x > 0$  and  $f'(x) < 0$  for  $x < 0$ . Hence  $f'$  changes from negative to positive at  $x_0 = 0$ , and since  $f$  is continuous at  $x_0 = 0$  this implies that  $f$  attains a local minimum at this point. Thus local extremes of a function may occur at places where the first derivative does *not* exist.

(b) An example is given by the function  $g(x) = x^3$ , which obviously satisfies  $g'(0) = 0$ , but the origin is neither a local minimum nor a local maximum of  $g$  (recall the graph of  $g$ , or see below). This shows that stationary points are not necessary local extremes.

(c) An example is given by the (equilateral) hyperbola  $y = h(x) = 1/x$  with  $x \in A := \mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, +\infty)$ . Indeed, for  $x_1, x_2 \in (0, +\infty)$  with  $x_1 < x_2$  we have  $\frac{1}{x_1} > \frac{1}{x_2}$ , i.e.,  $h(x_1) > h(x_2)$ , and thus  $y = h(x)$  is strictly decreasing on  $(0, +\infty)$ . Since  $h$  is an odd function,  $h(-x) = -h(x)$  for all  $x \in A$ , we immediately deduce that  $h$  is strictly decreasing on  $(-\infty, 0)$ , as well, and hence on its whole domain  $A$ . This also follows by the first derivative test:  $h'(x) = -1/x^2 < 0$  for all  $x \in A$ . Hence there are no extreme points, a conclusion that one can illustrate by sketching the graph of  $h$  ( $h$  is an odd function, hence its graph is symmetric with respect to the origin, see also the figure below). Moreover, we have  $\lim_{x \rightarrow 0^+} h(x) = +\infty$  and hence the  $y$ -axis ( $x = 0$ ) is a vertical asymptote of  $h$ . In addition, the  $x$ -axis ( $y = 0$ ) is a horizontal asymptote, since  $\lim_{x \rightarrow +\infty} h(x) = 0$ .

In summary,  $f(x)$  is smooth on the whole of  $\mathbb{R}$ . It is strictly positive everywhere except for  $x = 0$ . All its derivatives at this point are zero.

It cannot be analytic at  $x_0 = 0$ . The limit of the function at the improper points  $\pm\infty$  is 1, while all its derivatives converge quickly to zero.

**6.1.10. Smooth jump functions.** The smooth functions are very “elastic” — from a local behaviour around one point we cannot deduce anything at all about the global behavior of such functions. On the other hand, analytic functions are completely determined just by derivatives at one point. In particular they are completely determined by their behaviour on an arbitrarily small neighbourhood of a single point from their domain. In this sense, analytic functions are very “rigid”.

In particular, the smooth functions allow for joining different constant values on disjoint open intervals in a differentiable way.

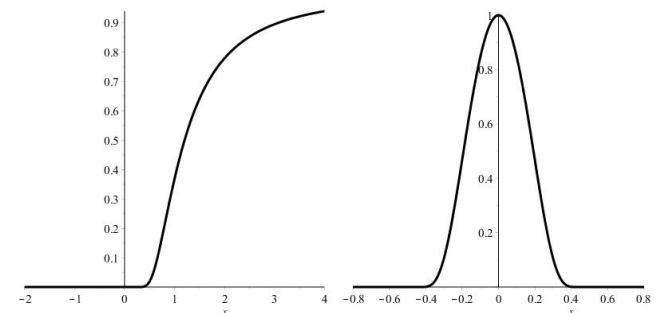
Let us look at such functions more closely now. We can modify  $f(x)$  from the previous paragraph in this way:

$$g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ e^{-1/x^2} & \text{if } x > 0. \end{cases}$$

Again it is a smooth function on all of  $\mathbb{R}$ . By another modification there is another function  $h$ , which is nonzero at all inner points of the interval  $[-a, a]$ ,  $a > 0$  and zero elsewhere.

$$h(x) = \begin{cases} 0 & \text{if } |x| \geq a \\ e^{\frac{1}{x^2-a^2} + \frac{1}{a^2}} & \text{if } |x| < a. \end{cases}$$

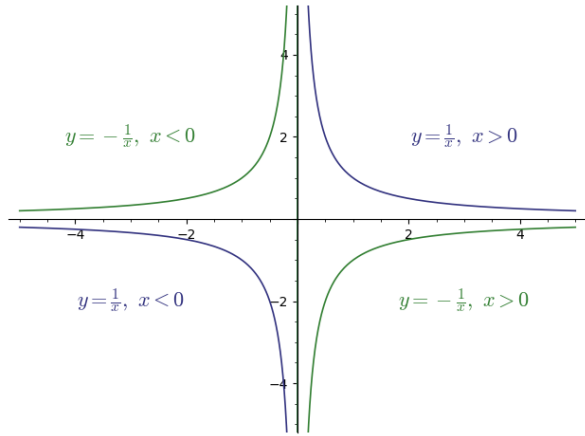
This function is again smooth on all of  $\mathbb{R}$ . The last two functions are in the two figures. On the right, the parameter  $a = 1/2$  is used.



Finally we show how to get smooth analogies of the Heaviside functions. For two fixed real numbers  $a < b$ , define the function  $f(x)$  exploiting the above function  $g$  as follows:

$$f(x) = \frac{g(x-a)}{g(x-a) + g(b-x)}.$$

For all  $x \in \mathbb{R}$  the denominator of the fraction is positive (because  $g$  is non-negative). For each of the three intervals determined by numbers  $a$  and  $b$  at least one of the summands of the denominator is nonzero). Thus the definition yields a smooth function  $f(x)$  on all of  $\mathbb{R}$ . For  $x \leq a$  the numerator of the fraction is zero according to the definition of  $g$ . For  $x \geq b$



Our figure also contains the graph of the function  $k(x) = -1/x$ , which is symmetric to the graph of  $h(x) = 1/x$  with respect to the  $x$ -axis. Notice that  $k$  has the same domain as  $h$  and is strictly increasing on  $A$ , since  $k'(x) = \frac{1}{x^2} > 0$  for all  $x \in A$ . Hence it provides another example of a function having no local minima or maxima. A generalization occurs by the hyperbolas  $y = a/x$ , with  $a > 0$  and  $a < 0$ , respectively.

For convenience, let us finally include the code used in Sage to construct these graphs:

```
h(x)=1/x; k(x)=-1/x
p=plot(h(x), x, -5, 5, ymin=-5, ymax=5,
  rgbcolor=(0.2,0.2,0.5), thickness=1.2)
p+=text(r"$y=\frac{1}{x}$", \ x>0$, (3, 2),
  fontsize=16, rgbcolor=(0.2,0.2,0.5))
p+=text(r"$y=\frac{1}{x}$", \ x<0$, (-3, -2),
  fontsize=16, rgbcolor=(0.2,0.2,0.5))
p+=plot(k(x), x, -5, 5, ymin=-5, ymax=5,
  rgbcolor=(0.2,0.5,0.2), thickness=1.2)
p+=text(r"$y=-\frac{1}{x}$", \ x<0$, (-3, 2),
  fontsize=16, rgbcolor=(0.2,0.5,0.2))
p+=text(r"$y=-\frac{1}{x}$", \ x>0$, (3, -2),
  fontsize=16, rgbcolor=(0.2,0.5,0.2)); show(p)
```

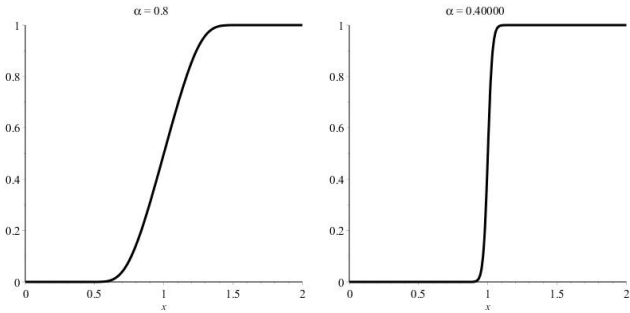
□

**6.A.21. An annoying technicality.** Generally speaking, Sage is still a work “in progress” and hence its users may face some technicalities. For instance, try to produce the graph of  $f(x) = x^{\frac{2}{3}}$  with  $x \in \mathbb{R}$  (or of  $x \mapsto x^{\frac{1}{3}} = \sqrt[3]{x}$ ), in a traditional way. In this case Sage will sketch only the portion corresponding to the positive reals. This problem occurs since Sage returns complex numbers for odd roots of negative numbers, when the latter are numerically approximated. This is the situation when, for example, one tries to sketch the graph of  $f$ . In order to realize this fact, in your editor type the syntax

```
c1 = (-1)^(2/3); print(c1)
print(float(c1))
c2 = (-1.)^(2/3); print(c2); float(c2)
```

For  $c1$  and by the third command, Sage returns the error “unable to simplify to float approximation”, while for  $c2$  the error

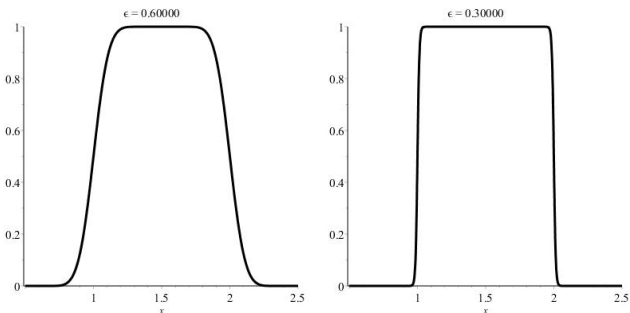
the numerator and denominator are equal. In the next two figures there are functions  $f(x)$  with parameters  $a = 1 - \alpha$ ,  $b = 1 + \alpha$ . On the left  $\alpha = 0.8$ , and on the right  $\alpha = 0.4$ .



Finally, we can create a smooth analogue of the characteristic function of any interval  $[c, d]$ .

Write  $f_\varepsilon(x)$  for the latter function  $f(x)$  with parameters  $a = -\varepsilon$ ,  $b = +\varepsilon$ . For the interval  $(c, d)$  with the length  $d - c > 2\varepsilon$  define the function  $h_\varepsilon(x) = f_\varepsilon(x - c) \cdot f_\varepsilon(d - x)$ . This function is identically zero on the intervals  $(-\infty, c - \varepsilon)$  and  $(d + \varepsilon, \infty)$ . It is identically one on the interval  $(c + \varepsilon, d - \varepsilon)$ . Moreover, it is smooth everywhere. Locally it is either constant or monotonic (you should verify the last claim yourself). The smaller the  $\varepsilon > 0$ , the faster  $h_\varepsilon(x)$  jumps from zero to one around the beginning of the interval or back at the end of it.

The diagram shows the choices  $[c, d] = [1, 2]$  and  $\varepsilon = 0.6$ ,  $\varepsilon = 0.3$ .



**6.1.11. Differential of a function.**

In practical use of differential calculus, we often work with dependencies between several variables, say  $y$  and  $x$ . The choice of dependent and independent variable is not fixed. The explicit relation  $y = f(x)$  with some function  $f$  is then only one of possible options. Differentiation then expresses that the immediate change of  $y = f(x)$  is proportional to the immediate change of  $x$  with the proportion of  $f'(x) = \frac{df}{dx}(x)$ .

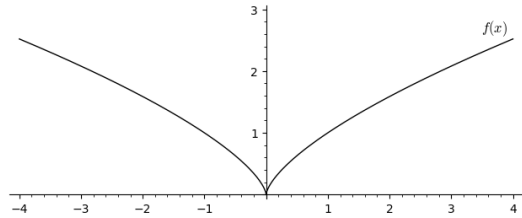
This relation is often written as  $dy = f'(x)dx$ , or

$$df(x) = f'(x)dx,$$

where we interpret  $df(x)$  as a linear map  $\mathbb{R} \rightarrow \mathbb{R}$  defined on increments of  $x$  at the given point,  $df(x)(v) = f'(x) \cdot v$ , while the identity map yields  $dx(x)(v) = v$ .

We talk about the *differential of function*  $f$ , which is the best linear approximation of the function  $f$  around the point

from the command `float(c2)` is about the difficulty of converting a complex number to float (notice the difference in the definition of `c1`, `c2`). In this case Sage advises us to use some of the options `abs()` or `real_part()`. Motivated by this, to cure our problem for the given  $f$  we may instead plot the graph of  $|x|^{\frac{2}{3}}$ , see the figure below.



**6.A.22.** Recall that if  $x_0$  is a critical point of a differentiable function and  $f''(x_0) > 0$ , then  $f'$  is strictly increasing near  $x_0$ . Therefore, the conditions  $f'(x_0) = 0$  and  $f''(x_0) > 0$  imply that  $x_0$  is a local minimum. Similarly, the conditions  $f'(x_0) = 0$  and  $f''(x_0) < 0$  imply that  $x_0$  is a local maximum. Show with a counterexample that the converse of these conclusions are not true in general.  $\circ$

**6.A.23.** Suppose that a function  $f : \mathbb{R} \rightarrow \mathbb{R}$  admits the line  $y = 3x + 1$  as an asymptote with slope, as  $x \rightarrow +\infty$ . Compute the limit

$$\lim_{x \rightarrow +\infty} \frac{xf(x) + 7x^2 + 2}{x^2f(x) + 3x^3 + 5x^2 + 2}.$$

**Solution.** By the definition of an asymptote with slope, we have

$$\lim_{x \rightarrow +\infty} \frac{f(x)}{x} = 3, \text{ and } \lim_{x \rightarrow +\infty} (f(x) - 3x) = 1.$$

Thus, if we denote by  $A$  the given limit, by dividing both the numerator and denominator by  $x^2$  we get

$$\begin{aligned} A &= \lim_{x \rightarrow +\infty} \frac{\frac{f(x)}{x} + 7 + \frac{2}{x^2}}{f(x) - 3x + 5 + \frac{2}{x^2}} \\ &= \frac{\lim_{x \rightarrow +\infty} \left(\frac{f(x)}{x}\right) + \lim_{x \rightarrow +\infty} \left(7 + \frac{2}{x^2}\right)}{\lim_{x \rightarrow +\infty} (f(x) - 3x) + \lim_{x \rightarrow +\infty} \left(5 + \frac{2}{x^2}\right)} = \frac{3 + 7}{1 + 5} = 5/3. \end{aligned}$$

$\square$

*Let us describe a problem which requires a bit of integration, but in a very elementary level and hence already treatable. Further similar problems we will analyze in Section B and also in the final section of this chapter.*

**6.A.24.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a twice differentiable function on  $\mathbb{R}$  satisfying for all  $x \in \mathbb{R}$  the relation

$$(\sqrt{2}x^2 + 1)f''(x) + 4\sqrt{2}xf'(x) + 2\sqrt{2}f(x) = 0. \quad (*)$$

(a) Find those differentiable functions  $g : \mathbb{R} \rightarrow \mathbb{R}$  satisfying  $g(x) = 2\sqrt{2}xf(x) + (\sqrt{2}x^2 + 1)f'(x)$ ;

(b) Based on your answer in (a), determine the function  $f$ , if in addition it is known that the graph  $C_f$  of  $f$  passes through

$x$ , i.e. the following approximation property is true:

$$(1) \quad \lim_{v \rightarrow 0} \frac{f(x+v) - f(x) - df(x)(v)}{v} = 0.$$

Since we are working in dimension 1, the definition is equivalent to the existence of the derivative  $f'(x)$  (all linear maps are just multiplications by a constant) and then the differential is  $df(x)(v) = f'(x)v$ . Later, we shall see that the situation is quite different for functions of more variables.

If the quantity  $x$  is expressed by another quantity  $t$ , e.g.  $x = g(t)$  and, moreover,  $g$  is differentiable, the chain rule for differentiating the composite functions says that  $f \circ g$  has the differential too and

$$df(t) = d(f \circ g)(t) = \frac{df}{dx}(x) \frac{dx}{dt}(t) dt.$$

We may also write  $dy = f'(x)dx = f'(g(t))g'(t)dt$ . Therefore  $dy$  can be seen as a linear approximation of the given quantity independently of the choice of the dependent variable.

**6.1.12. The numerical derivatives.** We shall conclude this



section with two straightforward applications of differentials. First we provide a brief introduction to the numerical procedures for differentiation. Then we discuss curves in the plane and space, starting with the graphs of functions. This will also provide first glimpses into the so called vector calculus (working with vector valued functions).

In the beginning of this textbook we discussed how to describe the values in a sequence if its immediate differences are known, (c.f. paragraphs 1.1.5, 1.2.1). Before proceeding the same way with the derivatives we clarify the connections between derivatives and differences. The key to this is the Taylor expansion with remainder.

Suppose that for some (sufficiently) differentiable function  $f(x)$  defined on the interval  $[a, b]$ , the values  $f_i = f(x_i)$  at the points  $x_0 = a, x_1, x_2, \dots, x_n = b$ , are given while  $x_i - x_{i-1} = h$  for some constant  $h > 0$  and all indices  $i = 1, \dots, n$ . Write the Taylor expansion of function  $f$  in the form

$$f(x_i \pm h) = f_i \pm hf'(x_i) + \frac{h^2}{2}f''(x_i) \pm \frac{h^3}{3!}f^{(3)}(x_i) + \dots$$

Suppose the expansion is terminated at the term containing  $h^k$  which is of order  $k$  in  $h$ . Then the actual error is bounded by

$$\frac{h^{k+1}}{(k+1)!} |f^{(k+1)}(x)|$$

on the interval  $[x_i - h, x_i + h]$ . If the  $(k+1)^{th}$  derivative  $f$  is continuous, it can be bounded by a constant. Then for small  $h$ , the error of the approximation by the Taylor polynomial of order  $k$  acts like  $h^{k+1}$  except for a constant multiple. Such an estimation is called an *asymptotic estimation*.

the origin and the tangent line of  $C_f$  at the origin is perpendicular to the line determined by the equation  $x + \sqrt{2}y - 1 = 0$ ;

(c) Use Sage to confirm that your answer in (b) satisfies indeed the initial condition (\*). Moreover, sketch the graph of  $f$  for  $-10 \leq x \leq 10$ . Where did we meet such a similar function earlier?

(d) Use Sage to specify the local extremes of  $f$  and characterize their type.

**Solution.** (a) By differentiating the relation  $g(x) = 2\sqrt{2}x f(x) + (\sqrt{2}x^2 + 1)f'(x)$ , with respect to  $x$ , we obtain

$$g'(x) = 2\sqrt{2}f(x) + 2\sqrt{2}x f'(x) + 2\sqrt{2}x f'(x) + (\sqrt{2}x^2 + 1)f''(x) = 0, \quad x \in \mathbb{R},$$

where the final equality follows by (\*). This implies that  $g(x) = c$  for all  $x \in \mathbb{R}$  for some constant  $c \in \mathbb{R}$ , i.e.,  $g$  is constant.

(b) One observes that

$$g(x) = 2\sqrt{2}x f(x) + (\sqrt{2}x^2 + 1)f'(x) = ((\sqrt{2}x^2 + 1)f(x))', \quad x \in \mathbb{R}.$$

In (a) we proved that  $g(x) = c$  for all  $x \in \mathbb{R}$ , hence a combination with the previous relation gives  $((\sqrt{2}x^2 + 1)f(x))' = c$ . Hence it is easy to guess that  $(\sqrt{2}x^2 + 1)f(x) = cx + \alpha$ , for some constant  $\alpha \in \mathbb{R}$ . As we will see in the forthcoming section, this follows by integrating both parts of the relation including the derivative, i.e.,

$$\int ((\sqrt{2}x^2 + 1)f(x))' dx = \int c dx \implies (\sqrt{2}x^2 + 1)f(x) = cx + \alpha, \quad \alpha \in \mathbb{R}.$$

The polynomial  $\sqrt{2}x^2 + 1$  has only complex roots, hence we can divide which yields the expression

$$f(x) = \frac{cx + \alpha}{\sqrt{2}x^2 + 1}, \quad \forall x \in \mathbb{R}. \quad (**)$$

To determine the constants  $c$  and  $\alpha$  we rely on the information specified by the scenario in (b). First,  $f$  passes through the origin, hence we need  $\alpha = 0$ . Next, the line  $x + \sqrt{2}y - 1 = 0$  is written as  $y = -\frac{1}{\sqrt{2}}x + 1 = -\frac{\sqrt{2}}{2}x + 1$ , which means that its slope equals to  $-\frac{\sqrt{2}}{2}$ . Since the tangent line of  $f$  at the origin should be perpendicular to this line, we need  $f'(0) = \frac{2}{\sqrt{2}} = \sqrt{2}$ , such that  $-\frac{\sqrt{2}}{2} \cdot \sqrt{2} = -1$ . Therefore, to determine  $c$  we need to solve the equation  $f'(0) = \sqrt{2}$ . So, let us use (\*\*) (for  $\alpha = 0$ ) to compute  $f'(x)$ :

$$f'(x) = \frac{c(-\sqrt{2}x^2 + 1)}{(\sqrt{2}x^2 + 1)^2}, \quad x \in \mathbb{R}.$$

Thus  $f'(0) = \sqrt{2}$  if and only if  $c = \sqrt{2}$ , which means that  $f(x) = \frac{\sqrt{2}x}{\sqrt{2}x^2 + 1}$ , for all  $x \in \mathbb{R}$ .

(c) A confirmation in Sage relies on the command `bool`, as follows (we also include the syntax for sketching  $C_f$ )

ASYMPTOTIC ESTIMATES

**Definition.** The expression  $G(h)$  is *asymptotically equal* to  $F(h)$  for  $h \rightarrow 0$ , write  $G(h) = O(F(h))$ , if the finite limit

$$\lim_{h \rightarrow 0} \frac{G(h)}{F(h)} = a \in \mathbb{R}$$

exists.

Similarly, compare the expressions for  $h \rightarrow \infty$  and use the same notation.

If the limit is zero, then we write  $G(h) = o(F(h))$ .

This way of denoting the asymptotic behaviour is often called the *big O* and *small o* notation. For instance, the differentiability of a function  $f$  means that the error of the approximation of  $f$  by its differential is  $o(|h|)$  for increments  $h$  of the argument.

Determine the values of the derivatives of  $f(x)$  at the points  $x_i$  as  $f_i^{(j)}$ . Write the Taylor expansion as:

$$f_{i \pm 1} = f_i \pm f'_i h + \frac{f''_i}{2} h^2 \pm \frac{f'''_i}{6} h^3 + \dots$$

Considering combinations of the two expansions and  $f_i$  itself, we can express the derivative  $f'_i$  as follows

$$\begin{aligned} \frac{f_{i+1} - f_{i-1}}{2h} &= f'_i + \frac{h^2}{3!} f_i^{(3)} + \dots \\ \frac{f_{i+1} - f_i}{h} &= f'_i + \frac{h}{2!} f''_i + \dots \\ \frac{f_i - f_{i-1}}{h} &= f'_i - \frac{h}{2!} f''_i + \dots \end{aligned}$$

This suggests a basic numerical approximation for derivatives:

CENTRAL, FORWARD, AND BACKWARD DIFFERENCES

The *central difference* is defined as

$$f'_i = \frac{f_{i+1} - f_{i-1}}{2h},$$

the *forward difference* is

$$f'_i = \frac{f_{i+1} - f_i}{h},$$

and the *backward difference* is

$$f'_i = \frac{f_i - f_{i-1}}{h}.$$

**Theorem.** The asymptotic estimate of the error of the central difference is  $O(h^2)$ . The errors of the backward and forward differences are  $O(h)$ .

**PROOF.** If we use the Taylor expansions with remainder of the appropriate order, we obtain an expression of the error of the approximation by the central difference in the form

$$\frac{1}{3!} h^2 (f^{(3)}(x_i + \xi h) - f^{(3)}(x_i - \eta h)).$$

```
f(x)=(sqrt(2)*x)/(sqrt(2)*x^2+1)
print(bool((sqrt(2)*x^2+1)*diff(f, x, 2) +
4*sqrt(2)*x*diff(f, x)+2*sqrt(2)*f(x)==0))
show(plot(f(x), x, -10, 10,
```

Here,  $0 \leq \xi, \eta \leq 1$  are the values from the remainder expression of  $f_{i+1}$  and  $f_{i-1}$ , respectively. The error of the second derivative in the other two cases is obtained similarly.  $\square$

Surprisingly, the central difference is one order better than the other two. But of course, the constants in the asymptotic estimates are important, too. In the case of the central difference, the bound on the third derivative appears, while in the two other cases second derivatives show up instead.

We proceed the same way when approximating the second derivative. To compute  $f''(x_i)$  from a suitable combination of the Taylor polynomials, we cancel both the first derivative and the value at  $x_i$ . The simplest combination cancels all the odd derivatives as well:

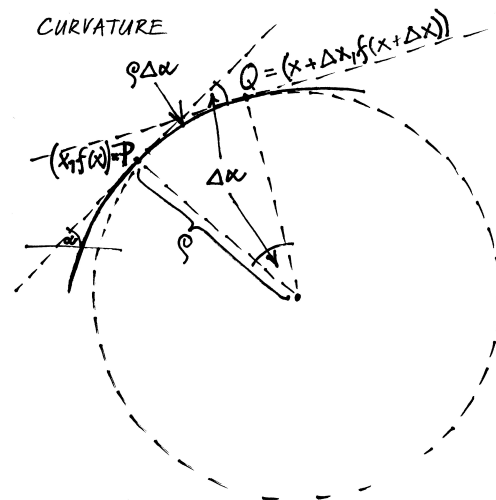
$$\frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} = f_i^{(2)} + \frac{h^2}{12}f^{(4)}(x_i) + \dots$$

This is called the *second order difference*. Just as in the central first order difference, the asymptotic estimate of the error is

$$f_i^{(2)} = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} + O(h^2).$$

Notice that the actual bound depends on the fourth derivative of  $f$ .

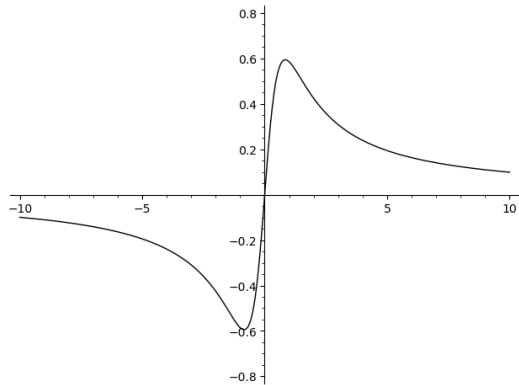
**6.1.13. The curvature of the graph of a function.** Imagine the graph of a function as a movement in the plane parametrized by the independent variable  $x$ . The vector  $(1, f'(x)) \in \mathbb{R}^2$  represents the velocity at  $x$  of such a movement. The tangent line through  $[x, f(x)]$  parametrized by this directional vector then represents a linear approximation of the curve. The goal is to discuss how “curved” is the graph at  $x$ . This is a straightforward exercise working with differentials in the setup of elementary plane geometry. It might need some effort to keep the attention.



If  $f''(x) = 0$  and simultaneously  $f'''(x) \neq 0$ , the graph of the function  $f$  intersects its tangent line. In such a case, the tangent line is the best approximation of the curve at the point  $x$  up to the second order as well. We describe this by saying that the graph of  $f$  has *zero curvature* at the point  $x$ .

```
ymin=-0.8, ymax=0.8, color="black"))
```

This block verifies our answer and produces the figure presented here:



(d) As shown in the graph above,  $f$  has two local extrema: one maximum and one minimum. Recall that critical points are found by solving the equation  $f'(x) = 0$ . Sage performs this calculation quickly for us:

```
f(x)=(sqrt(2)*x)/(sqrt(2)*x^2+1)
d1=diff(f, x).factor();
print(solve(d1==0, x))
```

This prints out two solutions, namely  $x_{\pm} = \pm \frac{1}{2}2^{3/4} = \pm \frac{\sqrt[4]{8}}{2} \approx \pm 0.84$ . An alternative to obtain these solutions is based on the command `roots` and can be implemented as follows:

```
f(x)=(sqrt(2)*x)/(sqrt(2)*x^2+1)
d1=diff(f, x).factor();
roots =d1.roots()
sols = []
for i in range(len(roots)):
    sols.append(roots[i][0])
print(sols)
```

Executing this block one gets the two solutions (without their multiplicities), i.e.,  $[-(1/2) * 2^{3/4}, (1/2) * 2^{3/4}]$ .

To verify that  $x_+$  (respectively,  $x_-$ ) is a point where  $f$  attains its local maximum (respectively, minimum) we will use the criterium of the second derivative. Hence it suffices to add the syntax

```
bool(diff(f, x, 2)(x=1/2*2^(3/4))<0)
bool(diff(f, x, 2)(x=-1/2*2^(3/4))>0)
```

which for both cases returns True. Or we can program Sage to check itself the critical points and print the corresponding result. For this it is sufficient to add in the initial block the following:

```
d2 = f.diff(2)
for a in sols:
    if d2(a)>0:
        print("{} is a local minimum"
              .format(a))
    if d2(a)<0:
        print("{} is a local maximum"
              .format(a))
```

The nonzero values of the first derivative describe the speed of the growth. Intuitively we expect the second derivative to describe the acceleration, including how “curved” the graph is. As a matter of convention, we want the curvature to be positive if the graph of the function is above its tangent.

The tangent at a fixed point  $P = [x, f(x)]$  is the limit of the secants. The lines passing through the points  $P$  and  $Q = [x + \Delta x, f(x + \Delta x)]$ . To approximate the second derivative, interpolate the points  $P$  and  $Q \neq P$  by the circle  $C_Q$ , whose center is at the intersection of the perpendicular lines to the tangents at  $P$  and  $Q$ .

It can be seen from the figure that if the angle between the tangent at the fixed point  $P$  and the  $x$  axis is  $\alpha$  and the angle between the tangent at the chosen point  $Q$  and the  $x$  axis is  $\alpha + \Delta\alpha$ , then the angle of the latter perpendicular lines is  $\Delta\alpha$  as well.

improve a bit the picture - should only  $\rho\Delta\alpha$  a perhaps make the measured arc thicker.

Denote the radius of the circle by  $\rho$ . Then the length of the arc between points  $P$  and  $Q$  is  $\rho\Delta\alpha$ . As  $Q$  approaches the fixed point  $P$ , the length  $\rho\Delta\alpha$  of the arc approaches the length  $\Delta s$  of the curve between  $P$  and  $Q$ , that is, the graph of the function  $f(x)$ . At the same time the circle approaches some limit circle  $C_P$ . This limit circle  $C_P$  is called the *osculating circle*.



Thus we arrive at the basic relation for the expected radius  $\rho$  of the circle  $C_P$  in terms of the linear approximations of the quantities:

$$\rho = \lim_{\Delta\alpha \rightarrow 0} \frac{\Delta s}{\Delta\alpha} = \frac{ds}{d\alpha}.$$

Notice that the quantity on the right hand side is well defined (independently of its rather intuitive justification).

Define the *curvature* of the graph of the function  $f$  at the point  $P$  as the number  $1/\rho$ . Zero curvature then corresponds to an infinite radius  $\rho$ .

For computing the radius  $\rho$  in terms of  $f$  we need to express the length of the arc  $s$  by the change of the angle  $\alpha$  and express the derivative of this function in terms of the derivative of  $f$ .

Notice that for an increasing angle  $\alpha$  the length of the arc can either increase or decrease, depending on whether the circle  $C_Q$  has its center above or below the graph of the function  $f$ . The sign of  $\rho$  then reflects whether the function is concave or convex. There is also the special case when the center “runs off” to infinity in the limit. Instead of a circle there is the tangent line.

There is no direct tool to compute the derivative  $\frac{ds}{d\alpha}$ . However,  $\text{tg } \alpha = \frac{df}{dx}$ . By differentiating this equality with respect to  $x$  we obtain (using the chain rule for differentials)

$$\frac{1}{(\cos \alpha)^2} \frac{d\alpha}{dx} = f''.$$

On the left hand side we can substitute

$$\frac{1}{(\cos \alpha)^2} = 1 + (\text{tg } \alpha)^2 = 1 + (f')^2$$



In this case the output looks like as follows:

```
-1/2*2^(3/4) is a local minimum
1/2*2^(3/4) is a local maximum
```

Finally, Sage can be used to compute the values of  $f$  at  $x_{\pm}$ , which allows us to summarize as follows: The point  $P_+ = [\frac{\sqrt[4]{8}}{2}, \frac{\sqrt[4]{2}}{2}]$  is a local maxima of  $f$ , while the point  $P_- = [-\frac{\sqrt[4]{8}}{2}, -\frac{\sqrt[4]{2}}{2}]$  is a local minima of  $f$ . Notice  $\pm \frac{\sqrt[4]{2}}{2} \approx \pm 0.59$ .  $\square$

**6.A.25. Local extrema numerically via Sage.** We should remark that Sage provides an in-built method which numerically finds the local extremes of a given function on an interval  $[a, b]$ , along with the points that these extremes values are attained. This technique relies on the commands `find_local_maximum(f, a, b)` and `find_local_minimum(f, a, b)`, respectively. Hence, for example, for the function  $f$  presented in 6.A.24 one could alternatively type

```
f(x)=(sqrt(2)*x)/(sqrt(2)*x^2+1);
show(find_local_maximum(f, -10, 10))
show(find_local_minimum(f, -10, 10))
```

This, indeed, returns the desired answers in the form  $(f(x_{\pm}), x_{\pm})$ , see here:

```
(0.5946035575013604, 0.8408963982254773),
(-0.5946035575013604, -0.8408963982254773).
```

**6.A.26.** Use Sage to sketch the real functions  $f, g, h, k$  for  $x \in [-2, 2]$ , and find numerically their local extrema, if any, where:

- (a)  $f(x) = \frac{1}{8}x^8 - \frac{3}{5}x^5 - 3x + 9$ ,
- (b)  $g(x) = x^2 - \frac{1}{6}x^3$ ,
- (c)  $h(x) = x\sqrt{x^2 + 4}$ ,
- (d)  $k(x) = \ln(x^3 + 8)$ .  $\circ$

**6.A.27.** Find all intervals on which the function  $f(x) = e^{-x^2}$ , with  $x \in \mathbb{R}$ , is concave.

**Solution.** We compute  $f'(x) = -2xe^{-x^2}$  and  $f''(x) = 2(2x^2 - 1)e^{-x^2}$  for all  $x \in \mathbb{R}$ . The sign of  $f''$  is given in the following table

$x$	$-\infty$	$-\sqrt{2}/2$	$\sqrt{2}/2$	$+\infty$
$f''(x)$	+	0	-	0
				+

Thus for any  $x \in (-\sqrt{2}/2, \sqrt{2}/2)$  we get  $f''(x) < 0$ , and the desired answer is given by the open interval  $(-\sqrt{2}/2, \sqrt{2}/2) \subset \mathbb{R}$ .  $\square$

**6.A.28. Remark.** Using Sage we might consider using the `bool` command, to verify the inequality  $f''(x) < 0$ , for all  $x \in (-\sqrt{2}/2, \sqrt{2}/2)$ , from the previous problem. Hence, for example, one could type

```
f(x)=exp(-x**2)
df2=diff(f, x, 2)
s=RealSet(-sqrt(2)/2, sqrt(2)/2)
bool(df2(x)<0 for x in s)
```

which implies (see the rule for differentiating inverse functions)

$$\frac{dx}{d\alpha} = \frac{1 + (\operatorname{tg} \alpha)^2}{f''} = \frac{1 + (f')^2}{f''}.$$

Now, we are almost finished, because the increment of the length of arc  $s$  dependent on  $x$  is given by the formula

$$\frac{ds}{dx} = (1 + (f')^2)^{1/2}.$$

Thus, by the chain rule,

$$\rho = \frac{ds}{d\alpha} = \frac{ds}{dx} \frac{dx}{d\alpha} = \frac{(1 + (f')^2)^{3/2}}{f''}.$$

Taking the reciprocal value, we have computed:

CURVATURE OF THE GRAPH

If  $f$  is in  $C^2$ , then the curvature of its graph is given by the formula

$$\rho^{-1}(x) = \frac{f''(x)}{(1 + (f'(x))^2)^{3/2}}$$

The result explains the relation between the curvature and the second derivative. The denominator of the fraction is always positive. It equals the third power of the length of the tangent vector of the given graph curve. The sign of the curvature is therefore given only by the sign of the second derivative, which confirms the ideas about concave and convex points of functions. In particular, in singular points, the curvature is just the second derivative.

If the second derivative is zero, the curvature  $1/\rho$  is also zero. If  $f''$  is large, then the radius  $\rho$  is small, thus the curvature is large as well.

Compute the curvature of simple functions yourself and use osculating circles while sketching their graphs. The computation at the critical points of the function  $f$  is easiest. The radius of the osculating circle is the reciprocal value of the second derivative with the corresponding sign.

**6.1.14. Vector differential calculus.** As mentioned already in the introduction to chapter five, most considerations related to differentiation are based on the fact that the functions are defined on real numbers and that their values can be added and multiplied by real numbers. That is why functions  $f : \mathbb{R} \rightarrow V$  need to have values in a vector space  $V$ . We call them *vector functions of one real variable* or more briefly *vector functions*.

Now, we digress to consider functions with values in the plane or in space. Thus,  $f : \mathbb{R} \rightarrow \mathbb{R}^2$  and  $f : \mathbb{R} \rightarrow \mathbb{R}^3$ . We consider (parametrized) curves in plane and space. We could work with values in  $\mathbb{R}^n$  for any finite dimension  $n$ .

For simplification, we work with the fixed standard bases  $e_i$  in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ . So curves are given by pairs or triples of real functions of one real variable, respectively. The vector function  $r$  in plane or space, respectively, is given by

$$r(t) = x(t)e_1 + y(t)e_2, \quad r(t) = x(t)e_1 + y(t)e_2 + z(t)e_3.$$

In this block we used the `RealSet` command to define the interval of interest. Although Sage returns `True`, this result can be misleading and we cannot trust this method. This is because if we replace the last line by `bool(df2(x) > 0 for x in s)`, Sage will still return `True`. In fact, the `bool(df2(x) < 0 for x in s)` command checks if the condition is true for some sampled points within the interval, but does not guarantee that the inequality holds for all points in the interval. This is because it only evaluates the condition at specific points and does not perform a rigorous check over the entire range. Therefore, one should exercise caution when interpreting results from computer algebra systems.

**6.A.29.** Show that the function  $f(x) = x^a$  with  $0 < a < 1$  is concave for all  $x > 0$ . ○

**6.A.30.** Use Sage to determine the asymptotes of the function  $f(x) = x/(x^2 - 1)$ , with  $x \in \mathbb{R} \setminus \{\pm 1\}$ , and then sketch these asymptotes along with the graph of  $f$ . ○

**6.A.31.** Consider the function  $f(x) = \frac{e^x - 1}{e^x + 1}$  with  $x \in \mathbb{R}$ .  
 (a) Prove that  $f$  is increasing on  $\mathbb{R}$ , find its asymptotes and determine its range. Then use Sage to sketch the asymptotes along with the graph of  $f$ .  
 (b) Determine the extremes points, inflection points and intervals of convexity of  $f$ , if any. ○

**6.A.32.** Study the local behaviour of the function  $f(x) = x/(x + 1)^2$ , with  $x \in A := \mathbb{R} \setminus \{-1\}$ .

**Solution.** Let us begin with the asymptotes of  $f$ . Obviously,  $f$  is continuous on its domain  $A$ . Thus, the only place we may have a vertical asymptote is at the endpoint  $-1$ . Indeed, we see that the line  $x = -1$  is a vertical asymptote since  $\lim_{x \rightarrow -1^+} f(x) = \lim_{x \rightarrow -1^-} f(x) = -\infty$ . We also have  $\lim_{x \rightarrow -\infty} f(x) = \lim_{x \rightarrow \infty} f(x) = 0$ , thus the line  $y = 0$  is a horizontal asymptote. In Sage a computation of these limits can be done as usual, i.e.,

```
f(x)=x/(x+1)^2
show(lim(f(x), x=-1, dir="left"))
show(lim(f(x), x=-1, dir="right"))
show(lim(f(x), x=-oo)); lim(f(x), x=oo)
```

Next, by adding the command `show(diff(f,x).factor())` we obtain the first derivative of  $f$ ,

$$f'(x) = -\frac{(x-1)}{(x+1)^3}, \quad x \in A. \quad (b)$$

Try to verify these results by hand. By (b) we deduce that the equation  $f'(x) = 0$  has a unique solution, given by  $x = 1$ , hence  $f$  has a unique critical point. There it appears a relative maximum of  $f$ , with value  $f(1) = 1/4$  (see below for an explanation).

To determine the critical points of  $f$  via Sage, add in the previous block the syntax

```
solve(diff(f, x)==0, x)
```

(the command `diff(f,x).roots()` also works, as we explained in 6.A.24). Now, by (b) we also deduce that

The derivative of such a vector function is a vector, which approximates the map  $r$  by a linear map of the real line to the plane or to the space.

In the plane it is

$$\frac{dr}{dt}(t) = r'(t) = x'(t)e_1 + y'(t)e_2$$

and similarly in space.

The differential of a vector function in this context is:

$$dr = \left( \frac{dx}{dt}e_1 + \frac{dy}{dt}e_2 + \frac{dz}{dt}e_3 \right) dt$$

where the expression on the right hand side is understood as “selecting” an increment of the scalar independent variable  $t$  and mapping it linearly by multiplying the vector of the three derivative components. Thus the corresponding increment of the vector quantity  $r$  is obtained (of course, only two components in the plane).

The notation  $r(t)$  is a convenient way to describe curves in space. For example  $r(t) = (a \cos t, a \sin t, bt)$  or  $r(t) = a \cos t e_1 + a \sin t e_2 + b e_3$  for fixed constants  $a, b$  describes a circular helix. Here the parameter  $t$  is related to a suitable angle measured around the  $z$ -axis. The derivative of  $r(t)$  at  $t = t_0$ , determines the direction of the tangent line at  $r(t_0)$ . In Newtonian mechanics, the parameter  $t$  can stand for time, measured in suitable units. In this case the derivative of  $r(t)$  at time  $t = t_0$ , gives the velocity vector at the same time. The second derivative then represents the acceleration vector at the same time.

**6.1.15. Differentiating composite maps.** In linear algebra and geometry there are very useful special maps called forms. They have one or more vectors as their arguments and they are linear in each of their arguments. In this way we defined the length of the vectors (the dot product is a symmetric bilinear form) or the volume of a parallelepiped (this is an  $n$ -linear antisymmetric form, where  $n$  is the dimension of the space), see for example the paragraphs 2.3.23 a 4.1.19.

Of course, we insert vectors  $r(t)$  dependent on a parameter as the arguments of these operations. By a straightforward usage of the Leibniz rule for differentiation of a product of functions, the following is verified:

**Theorem.** (1) If  $r(t) : \mathbb{R} \rightarrow \mathbb{R}^n$  is a differentiable vector function and  $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a linear map, then the derivative of the map  $\Psi \circ r$  satisfies

$$\frac{d(\Psi \circ r)}{dt} = \Psi \circ \frac{dr}{dt}.$$

(2) Consider differentiable vectors  $r_1, \dots, r_k : \mathbb{R} \rightarrow \mathbb{R}^n$  and a  $k$ -linear form  $\Phi : \mathbb{R}^n \times \dots \times \mathbb{R}^n \rightarrow \mathbb{R}$  on the space  $\mathbb{R}^n$ . Then the derivative of the composed map

$$\varphi(t) = \Phi(r_1(t), \dots, r_k(t))$$

satisfies the (generalized) Leibniz's rule

$$\frac{d\varphi}{dt} = \Phi\left(\frac{dr_1}{dt}, r_2, \dots, r_k\right) + \dots + \Phi\left(r_1, \dots, r_{k-1}, \frac{dr_k}{dt}\right).$$

- $f'(x) < 0$  for all  $x \in (-\infty, -1) \cup (1, +\infty)$ ;
- $f'(x) > 0$  for all  $x \in (-1, 1)$ .

Thus  $f$  decreases for all  $x \in (-\infty, -1) \cup (1, +\infty)$  and increases for  $x \in (-1, 1)$ .

Let us now proceed with the second derivative. By adding the command `diff(f, x, 2)` we obtain the second derivative of  $f$ , given by  $f''(x) = \frac{2(x-2)}{(x+1)^4}$  for all  $x \in A$ . Hence  $x = 2$  is the unique solution of  $f''(x) = 0$  and we see that

- $f''(x) < 0$  for all  $x \in (-\infty, -1) \cup (-1, 2)$ ;
- $f''(x) > 0$  for all  $x \in (2, +\infty)$ .

Thus,  $f$  is convex for all  $x \in (2, +\infty)$  and concave for  $x \in (-\infty, -1) \cup (-1, 2)$ . It also follows that the unique inflection point of  $f$  appears at  $x = 2$ , with value  $f(2) = 2/9$ . Notice in order to verify that at the critical point located above, a relative maximum appears, we may use the criterion based on the second derivative (see below) and show that  $f''(1) < 0$ .

Let us finally focus on the graph of  $f$ . The origin is the unique intersection of the graph of  $f$  with the axes, while  $f$  is neither even nor odd (so its graph has no particular symmetries). To sketch the graph of  $f$  for certain  $x \in A$ , together with its vertical asymptote  $x = -1$ , we use the cell

```
f(x)=x/(x+1)^2
pf=plot(f, x, -5, 5, ymin=-5, ymax=1, color="black")
pm = point((1, 1/4), size=30, color="red")
pt = circle((-1, 0), 0.05, color="black")
pnf = point((2, 2/9), size=30, color="blue")
sy = line([(-1, -5), (-1, 5)], linestyle="--", thickness=0.8)
(pf+sy+pm+pt+pnf).show(ticks=1, tick_formatter=1)
```

Execute this block in your editor, to revise its “small” particularities (as red and blue colours for the presentation of the maximum and the inflection point of  $f$ , respectively). □

We will now focus on derivatives of functions defined “implicitly” by equations of the form  $F(x, y) = 0$ . In such cases it can be often impossible to solve the equation with respect to  $y$  (as an explicit function of  $x$ ). However, as we will see below it is still possible to find the derivative  $dy/dx$ , compute extrema and/or decide for the convexity features of  $y$ , etc. For instance, to compute  $dy/dx$  the trick is just to differentiate both sides of the given equation, and then solve for the derivative we are seeking for. This is the so called implicit differentiation, which has many useful applications.

**6.A.33. Implicit differentiation.** Find the extreme points of the real function  $y = f(x)$  given in the implicit form  $xy^2 - x^2y = 54$ .

**Solution.** Since  $y$  is a function of  $x$  we can differentiate the given relation with respect to  $x$ . We get

$$y^2 + 2xyy' - 2xy - x^2y' = 0,$$

and thus solving with respect to  $y'$  one deduces that

$$y' = \frac{2xy - y^2}{2xy - x^2}. \quad (\#)$$

(3) The previous statement remains valid even if  $\Phi$  also has values in the vector space, and is linear in all its  $k$  arguments.

**PROOF.** (1) The linear maps are given by a constant matrix of scalars  $A = (a_{ij})$  so that

$$\Psi \circ r(t) = \left( \sum_{i=1}^n a_{1i}r_i(t), \dots, \sum_{i=1}^n a_{mi}r_i(t) \right).$$

Carry out the differentiation separately for individual coordinates of the result. However, the derivative acts linearly with respect to scalar linear combinations, see Theorem 5.3.4. That is why the derivative is obtained simply by evaluating the original linear map  $\Psi$  on the derivative  $r'(t)$ .

(2) The second statement is obtained analogously. Write out the evaluation of the  $k$ -linear form on the vectors  $r_1, \dots, r_k$  in the coordinates in this way:

$$\Phi(r_1(t), \dots, r_k(t)) = \sum_{i_1, \dots, i_k=1}^n B_{i_1 \dots i_k} \cdot (r_1)_{i_1}(t) \dots (r_k)_{i_k}(t),$$

where the scalars  $B_{i_1 \dots i_k} = \Phi(e_{i_1}, \dots, e_{i_k})$  are given as the values of the given form on the chosen  $k$ -tuple of base vectors for every choice of indices. The rule for differentiating a product of scalar functions then yields the statement.

(3) If  $\Phi$  has vector values, it is given by finitely many components and the previous result can be used for each of them. □

In the Euclidean space  $\mathbb{R}^3$ , the scalar product assigns a scalar to two vectors, There is also the vector product, which assigns the vector  $u \times v \in \mathbb{R}^3$  to vectors  $u$  and  $v$ , see 4.1.21. This vector  $u \times v$  is orthogonal to both vectors  $u$  and  $v$ , its length equals the area of the parallelogram determined by  $u$  and  $v$  (in this order) and the orientation is such that the triple  $u, v, u \times v$  is a positively oriented basis.

The previous ideas immediately imply:

**Corollary.** Consider the vectors  $u(t)$  and  $v(t)$  in the space  $\mathbb{R}^3$ . The derivatives of their scalar product  $\langle u(t), v(t) \rangle$  and their vector product  $u(t) \times v(t)$  satisfy

- (1)  $\frac{d}{dt} \langle u(t), v(t) \rangle = \langle u'(t), v(t) \rangle + \langle u(t), v'(t) \rangle$
- (2)  $\frac{d}{dt} (u(t) \times v(t)) = u'(t) \times v(t) + u(t) \times v'(t)$

**6.1.16. The curvature of curves.** We develop far more powerful tools for studying curves in a more systematic way than when we discussed the curvature of the graphs of functions. We proceed in dimension three. Plane curves are a special case in which the third component is the constant zero.

Let  $r(s)$  be a curve in the Euclidean space  $\mathbb{R}^3$ . For our purposes, it is convenient to choose arc length  $s$  as for the parameter. It follows that  $\|\frac{dr}{ds}\| = 1$ , so that the tangent vector has unit length. When  $s$  is the parameter, the notation  $'$  is used for differentiation.



One can obtain this formula also in Sage by the cell

```
var("x"); y = function("y")(x)
eq = x*y**2 - y*x**2 == 54
dy = diff(y, x) #declare the derivative
sol = solve(diff(eq), dy) #solve the implicit
show(dy.subs(sol)) #substitute the solution
```

Thus Sage can be quickly programmed to implement an implicit differentiation for us, see also below.

Let us now return back to our task. Based on (#), we see that the equation  $y' = 0$  is equivalent to  $y(2x - y) = 0$ , which gives  $y = 0$ , or  $y = 2x$ . The first solution is not acceptable and hence we have  $y' = 0$  if and only if  $y = 2x$ . For  $y = 2x$  the given equation reduces to  $x^3 - 3^3 = 0$ , that is,  $(x - 3)(x^2 + 3x + 9) = 0$ . The real  $x = 3$  is the unique real solution of this equation, hence we get a stationary point of  $y$  with value  $y(3) = 6$ .

Next we need to characterize this critical point, and hence we need the second derivative. A differentiation of both sides in (#) gives the relation  $y'' = A(x)/B(x)$ , where  $A(x) = 2[(y + xy' - yy')(2xy - x^2) - (2xy - y^2)(y + xy' - x)]$  and  $B(x) = (2xy - x^2)^2$ , respectively. By replacing the coordinates  $[x = 3, y = 6]$  and since  $y'$  vanishes at this point, we finally deduce that

$$y''(3) = 12/27 > 0.$$

Thus,  $x = 3$  is local minimum of  $y = f(x)$ , with value  $y_{\min} = 6$ .  $\square$

**6.A.34.** Suppose that  $f : \mathbb{R} \rightarrow \mathbb{R}$  is a function which is twice differentiable on  $\mathbb{R}$  and satisfies

$$f^2(x) - xf(x) + 2x^2 - 7 = 0, \quad x \in \mathbb{R}.$$

If  $a \in \mathbb{R}$  is a stationary (critical) point of  $f$ , prove that  $a = \pm \frac{\sqrt{2}}{2}$ . Does  $f$  admit some inflection point?  $\circ$



Recall by Chapter 5 that first order derivatives provide linear approximations, which can be used to estimate values of differentiable functions (cf. 5.C.18). This idea is revised in 6.1.11, in terms of the “differential” of a differentiable function  $y = f(x)$ , defined by  $df(x) = f'(x)dx$ , i.e.,  $dy = f'(x)dx$ . This is of course an equivalent way to interpret the differentiability of  $f$ , which is encoded by the relation  $f'(x) = \frac{dy}{dx}$ . Notice  $dx$  is an independent variable which we may assign any non-zero real number, while  $dy$  (or  $df(x)$ ) is necessarily a dependent variable.

The differential of  $f$  can be viewed as a linear form on  $\mathbb{R}$ , satisfying  $df(x)(h) = f'(x)h$ , for all  $h \in \mathbb{R}$ , and provides the best linear approximation of  $f$ , around  $x$ . Let us illustrate the most basic features of this concept via examples.

**6.A.35.** Compute the differentials of the functions  $f(x) = e^{x^2}$ ,  $g(x) = \ln(x^2 + 1)$  and  $k(x) = \arctan(e^x)$  with  $x \in \mathbb{R}$ . Next present an implementation of these differentials via Sage.  $\circ$

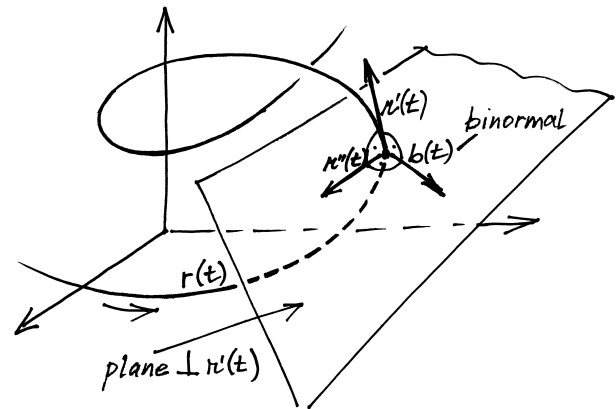
So  $\langle r'(s), r'(s) \rangle = 1$  for all  $s$ . The curve  $r(s)$  is parametrized by the length  $s$ . By another differentiation of this equality we arrive at (using the symmetry of the dot product)

$$0 = \frac{d}{ds} \langle r'(s), r'(s) \rangle = 2 \langle r''(s), r'(s) \rangle.$$

Thus the vector  $r''(s)$  is always orthogonal to the vector  $r'(s)$ .

This corresponds to the idea that after the choice of a parametrization with a derivative of constant length, the second derivative in the direction of the movement vanishes. The second derivative lies in the plane orthogonal to the tangent vector.

FRENET FRAME



If the second derivative is nonzero, the normed vector

$$n(s) = \frac{1}{\|r''(s)\|} r''(s)$$

is the (principal) normal of the curve  $r(s)$ . The scalar function  $\kappa(s)$  satisfying (at the points where  $r''(s) \neq 0$ )

$$r''(s) = \kappa(s)n(s)$$

is called the curvature of the curve  $r(s)$ . At the zero points of the second derivative  $\kappa(s)$  is defined as 0.

At the nonzero points of the curvature, the unit vector  $b(s) = r'(s) \times n(s)$  is well defined and is called the binormal of the curve  $r(s)$ . By direct computation

$$\begin{aligned} 0 &= \frac{d}{ds} \langle b(s), r'(s) \rangle = \langle b'(s), r'(s) \rangle + \langle b(s), r''(s) \rangle \\ &= \langle b'(s), r'(s) \rangle + \kappa(s) \langle b(s), n(s) \rangle = \langle b'(s), r'(s) \rangle, \end{aligned}$$

which shows that the derivative of the binormal is orthogonal to  $r'(s)$ . Further,  $b'(s)$  is also orthogonal to  $b(s)$  (for the same reason as with  $r'$  above). Therefore it is a multiple of the principal normal  $n(s)$ . We write

$$b'(s) = -\tau(s)n(s).$$

The scalar function  $\tau(s)$  is called the torsion of the curve  $r(s)$ .

In the case of plane curves, the definitions of binormal and torsion do not make sense.

**6.A.36. Differentials.** Construct in Sage a routine having as input a pair  $(f, x_0)$  consisting of a function and a point, and as output the differential of  $f$  at  $x_0$  (assuming that  $f$  is differentiable at  $x_0$ ). Next use your program to test some cases explicitly.



**Solution.** To solve this task we can use the `def` command to define a routine with the name `Dif`. Notice that `Dif` should have two inputs, a function `f` depending on one variable, and a real number `x_0`. Since  $df(x) = f'(x)dx$ , we should program Sage to view  $dx$  as a variable different than  $x$ , and it is sufficient to introduce  $dx$  as a symbolic variable, similarly with  $x$ . In total, one can program Sage via the block

```
var("x, dx")
def Dif(f, x_0) :
    f1=diff(f, x)(x=x_0)
    show("The differential of f(x)=", f, ", ", " ",
        "at x=", x_0, ", ", " ", "equals to", ":", " ", (dx)*f1)
    return
```

Notice in the second line of this block the syntax corresponds to the derivative  $f'(x_0)$ .

Let us now test the routine. For this goal, we chose the following pairs  $(f, x_0)$ :

```
Dif(ln(x), 2*pi)
Dif(exp(x), 0)
Dif(cos(2*x), 2*pi)
Dif(ln(x^2+1), 4*pi)
Dif(arctan(exp(x)-x^3+4*x), pi/2)
```

In the first case and for the pair  $(\ln(x), 2\pi)$ , Sage's output has the form:<sup>1</sup>

The differential of  $f(x)=\log(x)$ , at  $x=2\pi$ ,  
equals to:  $\frac{dx}{2\pi}$ ,

which is obviously true. Execute the rest commands in your editor and next verify Sage's solutions by hand.

As a side remark, observe that when our input is a non-differentiable function  $f$  at  $x_0$ , our routine will return an error (similarly with the error that Sage returns when the derivative of a function at a point does not exist).  $\square$

**6.A.37.** Consider a function  $y = f(x)$  differentiable at a point  $x_0$ . The change in the values of  $y$  in a small neighbourhood around  $x_0$ , is given by  $\Delta(y) = f(x_0 + dx) - f(x_0)$ . We can approximate linearly  $\Delta(y)$  using the differential of  $f$  at  $x_0$ , that is,  $\Delta(y) \approx f'(x_0)dx$ . Based on this formula compare  $\Delta(y)$  and  $dy$  for the cases:

- (a)  $y = x^4 - x^2 + 3x + 2$  and  $x$  changes from 2 to 2.05;
- (b)  $y = x^2 + \sin^2(x)\cos(x)$  and  $x$  changes from  $\frac{\pi}{2}$  to  $\frac{\pi}{2} + 0.04$ ;
- (c)  $y = \ln(x + 1) - e^{\sqrt{x}}$  and  $x$  changes from 1 to 1.03.

**Solution.** (a) We have  $x_0 = 2$  and  $x_0 + dx = 2 + 0.05 = 2.05$ . Thus  $\Delta(y) = f(x_0 + dx) - f(x_0) = f(2.05) - f(2) \approx$

<sup>1</sup>To avoid confusions, recall that in Sage the function `log(x)` corresponds to the natural logarithm  $\ln(x)$ .

We have not yet computed the rate of change of the principal normal, which can be written as  $n(s) = b(s) \times r'(s)$ :

$$\begin{aligned} n'(s) &= b'(s) \times r'(s) + \kappa(s)b(s) \times n(s) \\ &= -\tau(s)n(s) \times r'(s) + \kappa(s)(-r'(s)) \\ &= \tau(s)b(s) - \kappa(s)r'(s). \end{aligned}$$

Summarizing, for all points with nonzero second derivative of the curve  $r(s)$  parametrized by the arc length we constructed the nice orthonormal basis  $(r'(s), n(s), b(s))$ , called *the Frenet frame* in the classical literature. At the same time, this basis is used in order to express the derivatives of its components in the well known form

FRENET-SERRET FORMULAE

For the Frenet frame  $(r'(s), n(s), b(s))$ , the so called *Frenet-Serret formulae* hold true

$$\begin{aligned} \frac{dr'}{ds}(s) &= \kappa(s)n(s), & \frac{dn}{ds}(s) &= \tau(s)b(s) - \kappa(s)r'(s) \\ \frac{db}{ds}(s) &= -\tau(s)n(s). \end{aligned}$$

The following theorem tells how crucial the curvature and torsion are. Notice that if the curve  $r(s)$  lies in one plane, then the torsion is identically zero. In fact, the converse is true as well. We shall not provide the proofs here.

**Theorem.** Two curves in the space  $\mathbb{R}^3$  parametrized by the length of their arc can be mapped to each other by a Euclidean transformation (i.e. an affine map keeping the distances) if and only if their curvature functions and torsion functions coincide, except for a possible constant shift of the parameter. Moreover, for every choice of smooth functions  $\kappa$  and  $\tau$  there exists a smooth curve with these parameters.

By a straightforward computation we can check that the curvature of the graph of the function  $y = f(x)$  in plane and the curvature  $\kappa$  of this curve defined in this paragraph coincide. Indeed, comparing the differentials of the length of the arc for the graph of a function (as a curve with coordinates  $x(t), y(t) = f(x(t))$ ):

$$ds = (1 + (f_x)^2)^{1/2} dx, \quad dx = (1 + (f_x)^2)^{-1/2} ds$$

(we write  $f_x = \frac{df}{dx}$ ), we obtain the following equality for the unit tangent vector of the graph of a curve

$$r'(s) = ((1 + (f_x)^2)^{-1/2}, f_x(1 + (f_x)^2)^{-1/2}).$$

Now, we have to go through a bit messy computation for the second derivative. Let us write shortly  $r = (x, y)$ ,  $y' = f_x x'$ ,  $x' = (1 + f_x^2)^{-1/2}$  (remind  $'$  always means derivative with respect to the arc length parameter  $s$ ). Then

$$\begin{aligned} x'' &= -\frac{1}{2}(x')^3 2f_x f_{xx} x' = -(x')^4 f_x f_{xx} \\ y'' &= f_{xx}(x')^2 + f_x x'' = f_{xx}(x')^2 - f_{xx} f_x^2 (x')^4. \end{aligned}$$

1.6085. On the other hand,  $dy|_{x=x_0} = f'(x_0)dx$ , that is,

$$dy = (4x^3 - 2x + 3)|_{x=2} dx = 31 \cdot 0.05 = 1.55.$$

Thus  $|\Delta(y) - dy| = \Delta(y) - dy \approx 1.6085 - 1.55 \approx 0.0585$ .

(b) In this case we have  $x_0 = \pi/2$  and  $x_0 + dx = \frac{\pi}{2} + 0.04$ . Based on the following block in Sage,

```
f(x)=x^2+(sin(x))^2*cos(x)
D=f((pi/2)+0.04)-f(pi/2)
show(N(D))
```

we compute  $\Delta(y) = f(\frac{\pi}{2} + 0.04) - f(\frac{\pi}{2}) \approx 0.08733$ . Also, we compute that

$$f'(\frac{\pi}{2}) = (2x + 2\sin(x)\cos^2(x) - \sin^3(x))|_{x=\pi/2} = \pi - 1.$$

Hence for the differential we get  $dy|_{x=\pi/2} = (\pi - 1) \cdot 0.04 \approx 0.08566$ , and we deduce that  $|\Delta(y) - dy| = \Delta(y) - dy \approx 0.00167$ .

(c) Let us answer this task only with the aid of Sage, and leave the formal computation as an exercise. An appropriate syntax has the form

```
f(x)=ln(x+1)-e**(sqrt(x))
D=f(1.03)-f(1); show(N(f(1.03)-f(1)))
d=(diff(f(x),x)(x=1))*0.03; show(N(d))
show(N(abs(D-d)))
```

In this block the first `show` command implies that  $\Delta y \approx 0.025887$ , while the second gives  $dy \approx 0.025774$ . After introducing `f`, `D` and `d`, one could directly type the final command, which implies that  $|\Delta(y) - dy| \approx 0.00011$ .

At the end of the day, what we should keep from this task is that a formal computation of  $dy$  for all the three cases (a), (b) and (c), is easier than computing explicitly  $D(y)$ . In fact, for complicated functions the computation of  $D(y)$  can be extremely hard, but it is always easier to compute differentials. Finally, you may like to verify yourself that a receipt for improving the approximation is to choose smaller increments  $dx$ .  $\square$

**6.A.38.** Recall that the formula  $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$  determines the linear (tangential) approximation of a differentiable function  $f$  at a point  $x_0$ .

(a) Show that this is equivalent to the approximation formula  $\Delta(y) \approx dy$ ;

(b) For the function  $f(x) = 1/x$  compare the linear approximation of the value  $f(1.1)$ , obtained in 5.C.18, with the value determined by the relation  $f(x_0 + dx) \approx f(x) + dy$ .

**Solution.** (a) We have  $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$  and since  $x - x_0 = dx$ , this can be equivalently expressed as  $f(x) \approx f(x_0) + f'(x_0)dx$ . Using the definition of the differential of  $f$  at  $x_0$  (which we simply denote by  $dy$ ), we finally get the equivalent expression  $f(x) \approx f(x_0) + dy$ , that is,  $f(x_0 + dx) \approx f(x_0) + dy$ , where in the l.h.s the variable  $x$  was replaced by  $x_0 + dx$ . Hence, the tangential approximation is equivalent to the relation  $f(x_0 + dx) - f(x_0) \approx dy$ , i.e.,  $\Delta(y) \approx dy$ . This proves (a).

(b) Recall by 5.C.18 that the tangential approximation of  $1/x$  around  $x_0 = 1$  is the line  $L(x) = 2 - x$ , and hence

Hence

$$\begin{aligned} (x'')^2 + (y'')^2 &= f_{xx}^2 (x')^8 (f_x + (1 + f_x^2)^2 + f_x^4 \\ &\quad - 2f_x^2(1 + f_x^2)) \\ &= f_{xx}^2 (1 + f_x^2)^{-4} (f_x^2 + 1) \\ &= f_{xx}^2 (1 + f_x^2)^{-3}. \end{aligned}$$

We have arrived at the expected formula

$$\kappa^2 = \|r''\|^2 = \left(\frac{d^2f}{dx^2}\right)^2 (1 + (f_x)^2)^{-3}.$$

## 2. Integration

**6.2.1. Indefinite integral.** Now, we reverse the procedure of differentiation. We want to reconstruct the actual values of a function using its immediate changes. If we consider the given function  $f(x)$  as the (say continuous) derivative of an unknown function  $F(x)$ , then at the level of differentials we can write



$$dF = f(x) dx.$$

We call the function  $F$  the *primitive function* or the *indefinite integral* of the function  $f$  (we can also find the name *anti-derivative* in the literature). Traditionally we write

$$F(x) = \int f(x) dx.$$

**Lemma.** The primitive function  $F(x)$  to the function  $f(x)$  is determined uniquely on each interval  $[a, b]$  up to an additive constant.

**PROOF.** The statement follows immediately from Lagrange's mean value theorem, see 5.3.9. Indeed, if  $F'(x) = G'(x) = f(x)$  on the whole interval  $[a, b]$ , then the derivative of the function  $(F - G)(x)$  vanishes at all points  $c$  of the interval  $[a, b]$ . The mean value theorem implies that for all points  $x$  in this interval,

$$F(x) - G(x) = F(a) - G(a) + 0 \cdot (x - a).$$

Thus the difference of the values of the functions  $F$  and  $G$  is constant on the interval  $[a, b]$ .  $\square$

The previous lemma supports another notation for the indefinite integral:

$$F(x) = \int f(x) dx + C$$

with an unknown constant  $C$ .

The primitive functions are well defined for complex functions  $f$ , where the real and the imaginary part of the indefinite integrals are real primitive functions to the real and the imaginary parts of  $f$ . Thus, with no loss of generality, we work only with real functions in the sequel.

$L(1.1) = 0.9$ . Using the differential  $dy$  at  $x_0$ , by the claim in (a) we should get the same result. For, we have  $dy = -\frac{1}{x^2}dx$ . Moreover,  $1.1 = x_0 + dx = 1 + 0.1$ , i.e.,  $dx = 0.1$ , and hence we get  $dy|_{x_0=1} = -1 \cdot 0.1 = -0.1$ . Then the relation  $f(x_0 + dx) \approx f(x_0) + dy|_{x_0=1}$  gives

$$f(1.1) \approx f(1) + (-0.1) = 1 - 0.1 = 0.9,$$

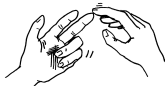
as required. □

The relation  $df(x)(h) = f'(x)h$  and the replacement of the increment  $dx$  by some small positive number  $h$  close to zero, allows us to rephrase the approximation formula  $\Delta(y) \approx dy$  as  $f(x+h) - f(x) \approx f'(x)h$ , i.e.,  $\frac{f(x+h) - f(x)}{h} \approx f'(x)$ . Based then on the Taylor expansion of  $f$  around  $x$ , and using the  $O$ -notation (see 6.1.12) we can finally arrive to the equation

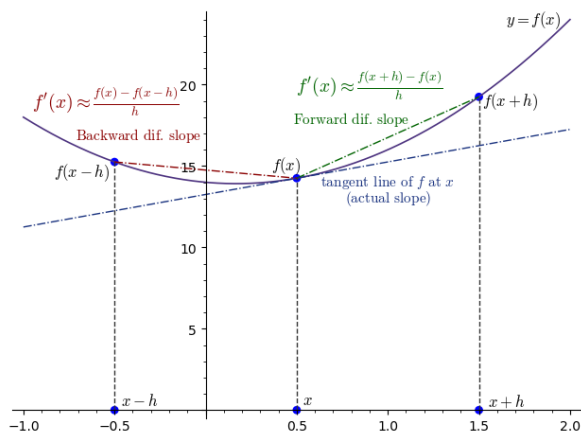
$$f'(x) = \frac{f(x+h) - f(x)}{h} + O(h).$$

Obviously, this establishes a first-order accurate approximation of  $f'(x)$ , since the dominate term in the truncation error is  $O(h)$ . The expression  $\frac{f(x+h) - f(x)}{h}$  is known as the “forward difference” formula for the first derivative. Similarly, one can consider the “backward difference” formula  $\frac{f(x) - f(x-h)}{h}$ , which gives another first-order accurate approximation of  $f'(x)$ , see 6.1.12.

**6.A.39. Remark on the role of  $h$ .**



The figure below illustrates the forward and backward difference formulas for the approximations of  $f'(0.5)$ , where  $f(x) = 3x^2 - x + 14$  with  $x \in I = [-1, 2]$  and  $h = 1$ . Check yourself that since  $h$  is big enough, this will not produce a “good” approximation of  $f'(0.5)$ . Hence, to improve the approximations in general small increments  $h$  closer to zero are more adaptable (see also below).



**6.A.40.** Consider the function  $f(x) = e^{x^2} - \frac{9}{10}$ . For the steps  $h = 0.1$  and  $h = 0.01$ , use Sage to find the forward and backward difference approximations of the derivative  $f'(0.2)$ .

**6.2.2. Newton integral.** We consider the value of a real function  $f(x)$  as an immediate increment of the region bounded by the graph of the function  $f$  and the  $x$  axis and try to find the area of this region between boundary values  $a$  and  $b$  of some interval. We relate this idea with the indefinite integral.

Suppose we are given a real function  $f$  and its indefinite integral  $F(x)$ , i.e.  $F'(x) = f(x)$  on the interval  $[a, b]$ .

Divide the interval  $[a, b]$  into  $n$  parts by choosing the points

$$a = x_0 < x_1 < \dots < x_n = b.$$

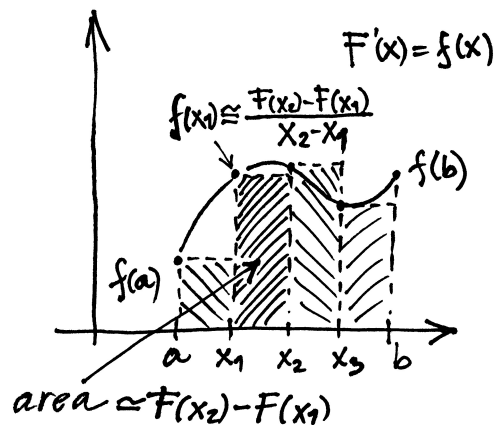
Approximate the values of the derivatives at the points  $x_i$  by the forward differences. That is, by the expressions

$$f(x_i) = F'(x_i) \simeq \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i}.$$

Finally the sum over all the intervals of our partition yields the approximation of the area:

$$\begin{aligned} \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i) &\simeq \sum_{i=0}^{n-1} \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i} (x_{i+1} - x_i) \\ &= \sum_{i=0}^{n-1} (F(x_{i+1}) - F(x_i)) = F(b) - F(a). \end{aligned}$$

Therefore we expect that for “nice enough” functions  $f(x)$ , the area of the region bounded by the graph of the function and the  $x$  axis (including the signs) can be calculated as a difference of the values of the primitive function at the boundary points of the interval. This procedure is called the *Newton integration*.<sup>3</sup>



<sup>3</sup>Isaac Newton (1642-1726) was a phenomenal English physicist and mathematician. The principles of integration and differentiation were formulated independently by him and Gottfried Leibniz in the late 17th century. It took nearly another two centuries before Bernhard Riemann introduced the completely rigorous modern version of the integration process.

Then, estimate the absolute value of the actual error. Which formula provides the better approximation and for which  $h$ ?

**Solution.** We have  $x = 0.2$ , hence the forward/backward difference formulas have the form

$$\frac{f(0.2 + h) - f(0.2)}{h}, \quad \frac{f(0.2) - f(0.2 - h)}{h}.$$

It is easy to estimate these expressions for the given  $h$  via Sage, and a recommendation is here:

```
f(x)=exp(x^(2))-0.9;
forw01=(f(0.2+0.1)-f(0.2))/0.1; show(forw01)
back01=(f(0.2)-f(0.2-0.1))/0.1; show(back01)
forw001=(f(0.2+0.01)-f(0.2))/0.01; show(forw001)
back001=(f(0.2)-f(0.2-0.01))/0.01; show(back001)
```

Based on Sage's output, for  $h = 0.1$ , we deduce that

$$f'(0.2) \stackrel{\text{forw. diff}}{\approx} \frac{f(0.3) - f(0.2)}{0.1} \approx 0.53364,$$

$$f'(0.2) \stackrel{\text{back. diff}}{\approx} \frac{f(0.2) - f(0.1)}{0.1} \approx 0.30761,$$

while for  $h = 0.01$  we get

$$f'(0.2) \stackrel{\text{forw. diff}}{\approx} \frac{f(0.21) - f(0.2)}{0.01} \approx 0.42761,$$

$$f'(0.2) \stackrel{\text{back. diff}}{\approx} \frac{f(0.2) - f(0.19)}{0.01} \approx 0.40513.$$

Of course, the second choice  $h = 0.01$  gives better approximations. For, the derivative of  $f$  is given by  $f'(x) = 2x e^{x^2}$ , thus  $f'(0.2) = 0.4163243$ . As for a confirmation via Sage, in the previous cell add the line

```
f1=diff(f(x), x)(x=0.2); show(f1)
```

Let us now estimate the errors, where it is reasonable to treat the case  $h = 0.01$  only, which provides smaller errors. So, to evaluate the differences  $|f'(0.2) - 0.42761|$ , and  $|f'(0.2) - 0.40513|$ , we may add in our block the lines

```
show(N(abs(f1-forw001)))
show(N(abs(f1-back001)))
```

We deduce that  $|f'(0.2) - 0.42761| \approx 0.0112841$  and  $|f'(0.2) - 0.40513| \approx 0.0111986$ . Thus, the backward difference with  $h = 0.01$  approximates  $f'(0.2)$  with the smallest actual error.  $\square$

**6.A.41.** For  $h = 0.1$ , and for the function  $f$  given in 6.A.40, use Sage to illustrate the forward and backward difference approximation of  $f'(0.2)$ , together with the tangent line of  $f$  at  $x = 0.2$ .  $\circ$

A combination of the forward and backward differences induces the "central difference" formula, given by

$$\frac{f(x + h) - f(x - h)}{2h}.$$

This provides another approximation of  $f'(x)$ , which has in fact truncation error being of order of  $O(h^2)$ , see 6.1.12.

**6.A.42.** For the function  $f$  introduced in 6.A.40, show that the central difference approximation of the derivative  $f'(0.2)$ ,

NEWTON INTEGRAL

If  $F$  is the primitive function to the function  $f$  on the interval  $[a, b]$ , then we write

$$\int_a^b f(x) dx = [F(x)]_a^b = F(b) - F(a)$$

and call it the *Newton (definite) integral* with the bounds  $a$  and  $b$ .

We prove later that for all continuous functions  $f \in C^0[a, b]$  the Newton integral exists and computes the area as expected. This is one of the fascinating theorems in elementary calculus. Before going into this, we discuss how to compute these integrals.

**6.2.3. Integration "by heart".** We show several procedures for computing the Newton integral. We exploit the knowledge of differentiation, and look for primitive functions.

The easiest case is the one where the given function is known as a derivative. To learn such cases, it suffices to read the tables for function derivatives in the menagerie the other way round. Hence:

INTEGRATION TABLE

For arbitrary nonzero  $a, b \in \mathbb{R}$  and  $n \in \mathbb{Z}, n \neq -1$ :

- $\int a dx = ax + C$
- $\int ax^n dx = \frac{a}{n+1} x^{n+1} + C$
- $\int e^{ax} dx = \frac{1}{a} e^{ax} + C$
- $\int \frac{a}{x} dx = a \ln x + C$
- $\int a \cos(bx) dx = \frac{a}{b} \sin(bx) + C$
- $\int a \sin(bx) dx = -\frac{a}{b} \cos(bx) + C$
- $\int a \cos(bx) \sin^n(bx) dx = \frac{a}{b(n+1)} \sin^{n+1}(bx) + C$
- $\int a \sin(bx) \cos^n(bx) dx = -\frac{a}{b(n+1)} \cos^{n+1}(bx) + C$
- $\int a \operatorname{tg}(bx) dx = -\frac{a}{b} \ln(\cos(bx)) + C$
- $\int \frac{a}{a^2 + x^2} dx = \operatorname{arctg}\left(\frac{x}{a}\right) + C$
- $\int \frac{-1}{\sqrt{a^2 - x^2}} dx = \arccos\left(\frac{x}{a}\right) + C$
- $\int \frac{1}{\sqrt{a^2 - x^2}} dx = \arcsin\left(\frac{x}{a}\right) + C.$

In all the above formulae, it is necessary to clarify the domain on which the indefinite integral is well defined. We leave this to the reader.



with step  $h = 0.01$ , is much better than the corresponding forward and backward difference approximations. In particular, show that this approximation coincides with  $f'(0.2)$ , up to the first four decimal digits. ○



*Calculus helps us understand various fundamental phenomena, such as the intrinsic geometry of curves and surfaces. We will start with a task that demonstrates the concept of the curvature of the graph of a function, as introduced in ???. To maintain continuity with our previous discussion (cf. 6.A.33), we recommend solving this task using implicit differentiation, though other methods are also available.*

**6.A.43.** Determine the curvature of the ellipse  $x^2 + 2y^2 = 2$  at its vertices. Also determine the equations of the circles of osculation at these vertices. ○



*Let us now explore how to work with parametrized curves in both plane and space within the context of calculus. These curves can be viewed as real-valued functions in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ , respectively, though the concept easily generalizes to  $\mathbb{R}^n$  (see 6.1.14 for a brief introduction to such functions, which are examples of “vector functions”).*

*These curves have numerous applications in Newtonian mechanics, where velocity and acceleration are described by the first and second derivatives of such parametric functions with respect to the parameter  $t$ , representing time. We will start with a simple task involving the derivatives of plane curves, which is left as an easy challenge for you and can be solved using Sage, as well.*

**6.A.44.** (a) Use the chain rule to prove that any differentiable plane curve  $\alpha(t) = [x(t), y(t)]$  satisfies

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt},$$

assuming that  $x'(t) = dx/dt \neq 0$ .

(b) Compute using two alternative methods the derivative  $\frac{dy}{dx}$  for the parametric curves

$$\alpha(t) = [t + 1, 2t^3], \quad \beta(t) = [1/(1 + t^2), \ln(t + 1)],$$

where  $t \in [0, 2]$  for both cases. Next evaluate  $\frac{dy}{dx}$  at  $t = 0$  and  $t = 2$  (if it is defined there).

(c) Use Sage to confirm your computations in (b) and then plot the curves  $\alpha, \beta$  via the command `parametric_plot`. ○

**6.A.45. Cycloid.** Consider the plane curve  $\alpha$  given by the parametric equations

$$x(t) = t - \sin(t), \quad y(t) = 1 - \cos(t), \quad t \in [0, 2\pi].$$

This is part of the so-called “cycloid”, which is the curve traced by a point on a circle as it rolls along a straight line

Further rules can be added by observations of special structure of the given functions. For example,

$$\int \frac{f'(x)}{f(x)} dx = \ln |f(x)| + C$$

for all continuously differentiable functions  $f$  on intervals where they are nonzero.

Of course, the rules for differentiating a sum of differentiable functions and constant multiples of differentiable functions yield analogous rules for the indefinite integral. So the sum of two indefinite integrals is the indefinite integral of the sum of the integrated functions, up to the freedom in the chosen constant, etc.

**6.2.4. Integration by parts.** The Leibniz rule for derivatives,  $(F \cdot G)'(t) = F'(t)G(t) + F(t)G'(t)$ , can be interpreted in the realm of the primitive functions. This observation leads to the following very useful practical procedure. It also has theoretical consequences.



INTEGRATION BY PARTS

The formula for computing the integral on the left hand side

$$\int F(x)G'(x) dx = F(x)G(x) - \int F'(x)G(x) dx + C$$

is called *integration by parts*.

The above formula is useful if we can compute  $G$  and at the same time compute the integral on the right hand side.

The principle is best shown on an example. Compute

$$I = \int x \sin x dx.$$

In this case the choice  $F(x) = x, G'(x) = \sin x$  will help. Then  $G(x) = -\cos x$  and therefore

$$I = x(-\cos x) - \int -\cos x dx = -x \cos x + \sin x + C.$$

Some integrals can be dealt with by inserting the factor 1, so that  $G'(x) = 1$ :

$$\begin{aligned} \int \ln x dx &= \int 1 \cdot \ln x dx \\ &= x \ln x - \int \frac{1}{x} x dx = x \ln x - x + C. \end{aligned}$$

**6.2.5. Integration by substitution.** Another useful procedure is derived from the chain rule for differentiating composite functions. If

$$F'(y) = f(y), \quad y = \varphi(x),$$

where  $\varphi$  is a differentiable function with nonzero derivative, then

$$\frac{dF(\varphi(x))}{dx} = F'(y) \cdot \varphi'(x)$$

and thus  $F(y) + C = \int f(y) dy$  can be computed as

$$F(\varphi(x)) + C = \int f(\varphi(x))\varphi'(x) dx.$$

without slipping.<sup>2</sup> Use calculus to show that the angle  $\theta$  of the tangent lines of  $\alpha$  at the points  $P = [x(\frac{2\pi}{3}), y(\frac{2\pi}{3})] \in \alpha$  and  $Q = [x(\frac{4\pi}{3}), y(\frac{4\pi}{3})] \in \alpha$ , respectively, is such that  $\tan(\theta) = -\sqrt{3}$ .

**Solution.** Recall that  $\sin(2\pi/3) = \sqrt{3}/2 = -\sin(4\pi/3)$  and  $\cos(2\pi/3) = -1/2 = \cos(4\pi/3)$ . Thus we compute  $P = [x_P, y_P] = [\frac{2\pi}{3} - \frac{\sqrt{3}}{2}, \frac{3}{2}]$  and  $Q = [x_Q, y_Q] = [\frac{4\pi}{3} + \frac{\sqrt{3}}{2}, \frac{3}{2}]$ , respectively. The tangent line of  $\alpha$  passing through  $P$  is the line

$$\ell_P(x) = y_P + k_{\frac{2\pi}{3}} \cdot (x - x_P), \quad k_{\frac{2\pi}{3}} = \left. \frac{dy}{dx} \right|_{t=\frac{2\pi}{3}}.$$

Similarly, the tangent line of  $\alpha$  passing through  $Q$  is the line

$$\ell_Q(x) = y_Q + k_{\frac{4\pi}{3}} \cdot (x - x_Q), \quad k_{\frac{4\pi}{3}} = \left. \frac{dy}{dx} \right|_{t=\frac{4\pi}{3}}.$$

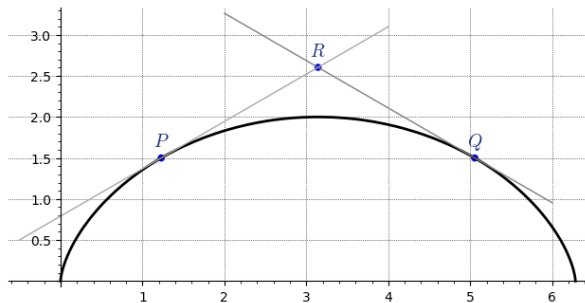
We compute

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{\sin(t)}{1 - \cos(t)} = \frac{1}{\tan(t/2)} = \cot(t/2),$$

and thus it follows that  $k_{\frac{2\pi}{3}} = \frac{\sqrt{3}}{3}$  and  $k_{\frac{4\pi}{3}} = -\frac{\sqrt{3}}{3}$ . Hence the angle  $\theta = \angle PRQ$ , where  $R$  is the intersection points of  $\ell_P, \ell_Q$ , is given by

$$\tan(\theta) = \frac{k_{\frac{4\pi}{3}} - k_{\frac{2\pi}{3}}}{1 + k_{\frac{2\pi}{3}} \cdot k_{\frac{4\pi}{3}}} = \frac{-\frac{\sqrt{3}}{3} - \frac{\sqrt{3}}{3}}{1 - \frac{1}{3}} = -\sqrt{3}.$$

An illustration is given here:



This figure and all the computations presented above can be done easily in Sage, and we leave this part as an exercise, see 6.A.46. For more details on the cycloid, see for example 6.D.17 in Section D.  $\square$

**6.A.46.** For the task in 6.A.45 illustrate the situation in Sage and confirm the related computations.  $\circ$

*In the end of the next section we will return back to parametric curves and learn how to compute their lengths and the areas of the regions bounded or enclosed by of their graphs. Hence soon we will also learn how to deal with integrals involving parametric equations.*

<sup>2</sup>When the circle has radius  $r$  the parametric equations of the cycloid are given by  $[r \cdot x(t), r \cdot y(t)]$ , and in our case we fixed  $r = 1$ . Be aware that the cycloid is *not* an ellipse.

By substituting  $x = \varphi^{-1}(y)$ , we obtain the originally desired primitive function. This is often written as follows:

#### INTEGRATION BY SUBSTITUTION

If  $\varphi(x)$  is differentiable with a nowhere vanishing derivative, then

$$F(y) = \int f(y) dy = \int f(\varphi(x))\varphi'(x) dx = F(\varphi(x)).$$

We talk about substituting the variable  $y$  by  $y = \varphi(x)$ .

On the level of differentials, the substitution can be easily understood in a way that (linearized) increments of the variable  $y$  and  $x$  are in mutual relation formally described by

$$dy = \varphi'(x) dx,$$

which corresponds the relation between the integrated quantities

$$f(y) dy = f(\varphi(x))\varphi'(x) dx.$$

As an illustration, we verify the last but one integral in the list in 6.2.3 using this method. To compute

$$I = \int \frac{1}{\sqrt{1-x^2}} dx.$$

Choose the substitution  $x = \sin t$ , for all  $t \in (-\frac{\pi}{2}, \frac{\pi}{2})$ . Then  $dx = \cos t dt$ . So

$$\begin{aligned} I &= \int \frac{1}{\sqrt{1-\sin^2 t}} \cos t dt = \int \frac{1}{\sqrt{\cos^2 t}} \cos t dt \\ &= \int dt = t + C. \end{aligned}$$

By substitution  $t = \arcsin x$  into the result,  $I = \arcsin x + C$ .

While substituting, the actual existence of the inverse function to  $y = \varphi(x)$  is required. To evaluate a definite Newton integral, it is needed to correctly recalculate the bounds of integration. Problems with the domains of the inverse functions can sometimes be avoided by dividing the integration into several intervals. We return to this point later.

**6.2.6. Integration by reduction to recurrences.** Often the use of substitutions and integrating by parts leads to recurrent relations, from which desired integrals can be evaluated. We illustrate by an example. Integrating by parts, to evaluate

$$\begin{aligned} I_m &= \int \cos^m x dx = \int \cos^{m-1} x \cos x dx \\ &= \cos^{m-1} x \sin x - (m-1) \int \cos^{m-2} x (-\sin x) \sin x dx \\ &= \cos^{m-1} x \sin x + (m-1) \int \cos^{m-2} x \sin^2 x dx. \end{aligned}$$

Using the formula  $\sin^2 x = 1 - \cos^2 x$ ,

$$mI_m = \cos^{m-1} x \sin x + (m-1)I_{m-2}.$$

The initial values are

$$I_0 = x, \quad I_1 = \sin x.$$

**B. Integration**

*In this section we will focus on tasks related to the notion of integration. As we pronounced before, the processes of differentiation and integration are inverse to each other, a result that we will confirm later in terms of the “fundamental theorem of calculus”. This theorem is in a sense the main link between the so called “differentiable calculus” and “integral calculus” and we will have the chances to analyze many examples a few below.*



*Let us begin with the notion of indefinite integrals, also called antiderivatives, which are introduced in 6.2.1 (see also 6.2.2).<sup>3</sup> We first present a few easy examples, based on basic rules from integration (see the “Integration Table” in 6.2.3).*

**6.B.1. Integration by heart.** Using integration “by heart” (see 6.2.3) evaluate the indefinite integrals given below. Next confirm your answers via Sage.

- (1)  $\int e^{-x} dx$ , with  $x \in \mathbb{R}$ ; (2)  $\int \frac{1}{\sqrt{4-x^2}} dx$ , with  $x \in (-2, 2)$ ; (3)  $\int \frac{1}{x^2+3} dx$ , with  $x \in \mathbb{R}$ ; (4)  $\int \frac{3x^2+1}{x^3+x+2} dx$ , with  $x \neq -1$ ; (5)  $\int |x| dx$ , with  $x \in \mathbb{R}$ .

**Solution.** (1) The primitive of  $f(x) = e^{-x}$  is the function  $F(x) = -e^{-x}$ , since  $(-e^{-x})' = e^{-x}$ . Thus, by the definition of the indefinite integral given in 6.2.1, we have  $\int f(x) dx = F(x) + C = -e^{-x} + C$ , for some constant  $C$ . (2) Use the formula  $\int \frac{dx}{\sqrt{1-x^2}} = \arcsin(x) + C$ , for some constant  $C \in \mathbb{R}$ . This gives

$$\int \frac{1}{\sqrt{4-x^2}} dx = \int \frac{\frac{1}{2}}{\sqrt{1-\left(\frac{x}{2}\right)^2}} dx = \arcsin\left(\frac{x}{2}\right) + C.$$

(3) Use the formula  $\int \frac{dx}{1+x^2} = \arctan(x) + C$ , for some constant  $C \in \mathbb{R}$ . We have

$$\begin{aligned} \int \frac{1}{x^2+3} dx &= \frac{1}{3} \int \frac{1}{\frac{x^2}{3}+1} dx = \frac{1}{\sqrt{3}} \int \frac{\frac{1}{\sqrt{3}}}{1+\left(\frac{x}{\sqrt{3}}\right)^2} dx \\ &= \frac{1}{\sqrt{3}} \arctan\left(\frac{x}{\sqrt{3}}\right) + C. \end{aligned}$$

(4) Use the formula  $\int \frac{f'(x)}{f(x)} dx = \ln|f(x)| + C$ , with  $C \in \mathbb{R}$ . Hence, we obtain

$$\int \frac{3x^2+3}{x^3+3x+2} dx = \ln|x^3+3x+2| + C.$$

(5) The function  $f(x) = |x|$  is continuous on  $\mathbb{R}$ , hence it has a primitive function on  $\mathbb{R}$ . This has the form  $\frac{x^2}{2} + c_1$ , for  $x \geq 0$ ,

<sup>3</sup>The notation for the indefinite integral was first introduced by Leibniz in 1675.

Integrals in which the integrated function depends on expressions of the form  $(x^2 + 1)$  can be reduced to these types of integrals using the substitution  $x = \operatorname{tg} t$ . For example, to compute

$$J_k = \int \frac{dx}{(x^2+1)^k}$$

the latter substitution yields (notice that  $dx = \cos^{-2} t dt$ )

$$J_k = \int \frac{dt}{\cos^2 t \left(\frac{\sin^2 t}{\cos^2 t} + 1\right)^k} = \int \cos^{2k-2} t dt.$$

For  $k = 2$ , the result is

$$J_2 = \frac{1}{2}(\cos t \sin t + t) = \frac{1}{2} \left( \frac{\operatorname{tg} t}{1 + \operatorname{tg}^2 t} + t \right).$$

After the reverse substitution  $t = \operatorname{arctg} x$

$$J_2 = \frac{1}{2} \left( \frac{x}{1+x^2} + \operatorname{arctg} x \right) + C.$$

When evaluating definite integrals, we can compute the whole recurrence after evaluating with the given bounds. For example while integrating over the interval  $[0, 2\pi]$ , the integrals have these values:

$$I_0 = \int_0^{2\pi} dx = [x]_0^{2\pi} = 2\pi$$

$$I_1 = \int_0^{2\pi} \cos x dx = [\sin x]_0^{2\pi} = 0$$

$$I_m = \int_0^{2\pi} \cos^m x dx = \begin{cases} 0 & \text{for odd } m \\ \frac{m-1}{m} I_{m-2} & \text{for even } m. \end{cases}$$

Thus for even  $m = 2n$ , the result is

$$\int_0^{2\pi} \cos^{2n} x dx = \frac{(2n-1)(2n-3)\dots 3 \cdot 1}{2n(2n-2)\dots 2} 2\pi.$$

For odd  $m$  it is zero (as could be guessed from the graph of the function  $\cos x$ ).

**6.2.7. Integration of rational functions.** The next goal is the integration of the quotients of two polynomials  $f(x)/g(x)$ . There are several simplifications to start with.



If the degree of the polynomial  $f$  in the numerator is greater or equal to the degree of the polynomial  $g$  in the denominator, carry out the division with remainder (see the paragraph 5.1.2). This reduces the integration to a sum of two integrals.

The division provides

$$f = q \cdot g + h, \quad \frac{f}{g} = q + \frac{h}{g}.$$

Thus,  $\int f(x)/g(x) dx = \int q dx + \int h(x)/g(x) dx$  where the first integral is easy and the second one is again an expression of the type  $h(x)/g(x)$ , but with degree of  $g(x)$  strictly larger than the degree of  $h(x)$  (such functions are called *proper rational functions*).

and  $-\frac{x^2}{2} + c_2$ , for  $x \leq 0$ , for some constants  $c_1, c_2 \in \mathbb{R}$ . It is easy to see that  $c_1 = c_2 = c \in \mathbb{R}$  and hence

$$\int |x| dx = \frac{1}{2}x|x| + c = \begin{cases} \frac{x^2}{2} + c, & \text{if } x \geq 0; \\ -\frac{x^2}{2} + c, & \text{if } x \leq 0; \end{cases}$$

A powerful capability of Sage, as most of the available computer algebra programs, is its ability to integrate symbolically. In Sage we will learn to integrate functions in many different ways and also numerically. For antiderivatives (indefinite integrals) one uses the `integral` function, via the syntax `integral(f(x), x)`, which can be rewritten as `f.integral(x)`. An alternative reads as `f.integrate(x)`. Later we will see that the situation of definite integrals does not differ much.

Recall that the indefinite integral (as an infinite set) can be represented by one specific function and its translations. Hence, one could expect that Sage will print out a function with a constant  $C$ , as in the relation  $F(x) = \int f(x) dx + C$ . However, Sage ignores  $C$  and hence you should always assume that this is implicitly part of the answer. Keeping in mind these details, we are now ready to solve our task. This can be done by the cell

```
show(integral(e^(-x), x))
show(integral(1/(sqrt(4-x^2)), x))
show(integral(1/(x^2+3), x))
show(integral((3*x^2+1)/(x^3+x+2), x))
show(integral(abs(x)))
```

Check yourself that this provides the desired answers.

Often, computing integrals via Sage we may need to add restrictions, via the command `assume`. For instance, type

```
m=var("m"); assume(m>1)
show(integral(1/x^m, x))
```

Notice without the command `assume(m > 1)`, Sage is not able to produce a result. We will meet further such examples in the sequel.  $\square$

**6.B.2.** Let  $f$  be a continuous function on  $\mathbb{R}$  with  $f(x) \neq 0$ , for all  $x \in \mathbb{R}$ . Suppose that the primitive function  $F$  of  $f$  satisfies  $4F(x) = f(x)$ , for all  $x \in \mathbb{R}$ , and moreover that  $f(4) = 4$ . Find the type of  $f$ .  $\circ$

**6.B.3.** (a) Write a routine in Sage which will print out a primitive of an integrable function  $f$ .



(b) Use your routine to find the primitive of some of the functions presented in the “Integration Table” in 6.2.3.

**Solution.** This is an easy task that one can implement as follows:

```
function("f")(x)
def Primitive_function(f) :
    F(x)=integral(f, x).factor()
    show("A primitive function of f(x)=", f,
        " is given by ", ":", F(x))
    return
```

Thus we can assume that the degree of  $g$  is strictly larger than the degree of  $f$ . We introduce the procedure to integrate proper rational functions by a simple example.

Observe that we can integrate  $(a+x)^{-n}$ ,  $n > 1$ , and

$$\int \frac{1}{a+x} dx = \ln|a+x| + C.$$

Summing such simple fractions yields more complicated ones:

$$\frac{-2}{x+1} + \frac{6}{x+2} = \frac{4x+2}{x^2+3x+2}$$

which can be integrated directly:

$$\int \frac{4x+2}{x^2+3x+2} dx = -2 \ln|x+1| + 6 \ln|x+2| + C.$$

This suggests looking for a procedure to express proper rational functions as a sum of simple ones. In the example, it is straightforward to compute the unknown coefficients  $A$  and  $B$ , once the roots of the denominator are known:

$$\frac{4x+2}{x^2+3x+2} = \frac{4x+2}{(x+1)(x+2)} = \frac{A}{x+1} + \frac{B}{x+2}.$$

Multiply both sides by the polynomial  $x^2+3x+2$  from the denominator and compare coefficients of the individual powers of  $x$  in the resulting polynomials:

$$4x+2 = A(x+2)+B(x+1) \implies 2A+B = 2, A+B = 4.$$

This procedure is called *decomposition into partial fractions*. It is a purely algebraic procedure based on properties of polynomials.

Without loss of generality, suppose that the denominator  $g(x)$  and the numerator  $f(x)$  do not share any real or complex roots and that  $g(x)$  has exactly  $n$  distinct real roots  $a_1, \dots, a_n$ . Then the points  $a_1, \dots, a_n$  are all the discontinuities of the function  $f(x)/g(x)$ .

Split the expression  $\frac{f(x)}{g(x)}$  according to the factors of the denominator. Thus, assume  $g(x)$  is the product

$$g(x) = p(x)q(x)$$

of two coprime polynomials. By the Bezout identity (see 12.3.8 on the page 1082), which is a corollary of the polynomial division with a remainder, there exist polynomials  $a(x)$  and  $b(x)$  of degrees strictly less than the degree of  $g$  such that

$$a(x)p(x) + b(x)q(x) = 1.$$

Multiplying this equality by the quotient  $f(x)/g(x)$ , gives

$$\frac{f(x)}{g(x)} = \frac{a(x)}{q(x)} + \frac{b(x)}{p(x)}.$$

Thus, we may restrict our attention to cases where the denominator  $g(x)$  cannot be decomposed further into two coprime polynomials.

Suppose that the polynomial  $g(x)$  has only real roots. Then there is a unique decomposition into factors  $(x - a_i)^{n_i}$ , where  $n_i$  are the multiplicities of the roots  $a_i$ ,  $i = 1, \dots, k$ . By a sequential use of the latter procedure with coprime

(b) To test the routine it is easy:

```
a,b=var("a, b")
Primitive_function(a*x^3)
Primitive_function(e^(a*x))
Primitive_function(a/x)
Primitive_function(a*sin(x*b))
Primitive_function(a/(a^2+x^2))
Primitive_function(1/sqrt(a^2+x^2))
Primitive_function(x*abs(x))
Primitive_function((3*x^2)*sqrt(1+(1/x^2)))
```

For instance, for the final case Sage’s output has the form:

A primitive function of  $f(x) = 3x^2\sqrt{1 + \frac{1}{x^2}}$  is  
 given by  $x^3 \left(\frac{x^2+1}{x^2}\right)^{\frac{3}{2}}$

Observe that when the input  $f$  is an integrable function, but there is no function, built up of addition, subtraction, multiplication, division, roots, exponents, logarithms, trigonometric functions, and inverse trigonometric functions which will have, as its derivative the function  $f$ , then Sage will *not* provide an answer. Such are the functions

$$e^{-x^3} \sin(x), e^{-x^3} \sin(x^2),$$

or the “Gaussian” (also called the “*error function*”) that we will meet later (this has a crucial role in statistics) However, one can still numerically compute such integrals, a situation that we will encounter later.  $\square$

*Trigonometric identities are often useful when integrating expression involving trigonometric functions. Test your skills on trigonometric identities by evaluating the following indefinite integral.*

**6.B.4.** Compute the integral  $A = \int \sin^2(x) \cos^2(x) dx$ .  $\circ$

Next we present a simple application of antiderivatives, where a given “initial condition” allows us to compute explicitly the constant of integration  $C$ . In particular, this provides an example of solving a first order differential equation with an initial condition (also referred to as an “initial value problem”). Further such problems we will meet in the end of this section (see 6.B.64, 6.B.65).

**6.B.5. An initial value problem.** Determine a curve  $y = f(x)$  passing from the point  $P = [1, 3]$  with slope  $\frac{2}{5}x$ . Then use the command `dsolve` in Sage to confirm your answer.

**Solution.** By assumption the slope of  $y = f(x)$  is  $\frac{2}{5}x$ , which means that  $f'(x) = \frac{df}{dx} = \frac{2}{5}x$ . By integration, we get

$$f(x) = \int f'(x) dx = \int \frac{2}{5}x dx = \frac{2}{5} \int x dx = \frac{x^2}{5} + C.$$

To specify the constant  $C \in \mathbb{R}$  use the equation  $f(1) = 3$ , i.e.,  $\frac{1}{5} + C = 3$ . This gives  $C = 14/5$  and hence  $y = f(x) = \frac{1}{5}(x^2 + 14)$ .

polynomials  $p(x)$  and  $q(x)$ , we obtain a representation of  $f(x)/g(x)$  as a sum of fractions of the form

$$\frac{f(x)}{g(x)} = \frac{r_1(x)}{(x - a_1)^{n_1}} + \dots + \frac{r_k(x)}{(x - a_k)^{n_k}},$$

where the degrees of the polynomials  $r_i(x)$  are strictly smaller than the degrees of the denominators. Finally, each of the summands can be represented as a sum

$$\frac{r(x)}{(x - a)^n} = \frac{A_1}{x - a} + \frac{A_2}{(x - a)^2} + \dots + \frac{A_n}{(x - a)^n}.$$

Indeed, we multiply the equation by  $(x - a)^n$  and start comparing the coefficients from the highest powers of the polynomial  $r(x)$  and compute sequentially  $A_1, A_2, \dots$  after expanding all the products. This can be done faster by suitable additions and subtractions, starting by the highest orders. For example,

$$\frac{5x - 16}{(x - 2)^2} = 5 \frac{x - 2}{(x - 2)^2} - 6 \frac{1}{(x - 2)^2} = \frac{5}{x - 2} + \frac{-6}{(x - 2)^2}.$$

Finally, we have to handle the case, where there are not enough real roots. There always exists a factorization of  $g(x)$  into linear factors with complex roots (see the fundamental theorem of Algebra in 1.A.6 on page 6). The non-real roots always appear in conjugated pairs, since  $\overline{g(z)} = g(\bar{z})$  for a polynomial with real coefficients.

Repeating the above procedure for ratios of complex polynomials gives the same result, but with complex coefficients. If we insist in having real expressions only, we may collect the conjugate pairs together and get quadratic factors expressed as sums of squares  $(x - a)^2 + b^2$  and their powers. The procedure works well and guarantees that it is possible to find summands in the form of

$$\frac{Bx + C}{((x - a)^2 + b^2)^n}.$$

As in the real roots case, there is always a corresponding decomposition into partial fractions of the form

$$\frac{A_1x + B_1}{(x - a)^2 + b^2} + \dots + \frac{A_nx + B_n}{((x - a)^2 + b^2)^n}$$

in the case of a power  $((x - a)^2 + b^2)^n$  of such quadratic (irreducible) factor as well.

The factorization of the polynomials and the further computations might be quite time consuming. The reader could prefer to experiment with computer algebra software instead. This works well in Maple by calling the procedure `convert(h, parfrac, x)` that decomposes the expression  $h$  rationally dependent on the variable  $x$  into partial fractions.

The important point is that we can already integrate all of the above partial fractions. The last mentioned ones lead to integrals discussed in example 6.2.6.

In summary, the rational functions  $f(x)/g(x)$  can be integrated easily, if the corresponding decomposition of the polynomial in the denominator  $g(x)$  is known. The reality is not that simple when computing (definite) Newton integrals. Although we find the primitive functions, the problematic points are the discontinuities of rational functions, in whose

To solve the system  $\{\frac{df}{dx} = \frac{2}{5}x, f(1) = 3\}$  with Sage we will use the command `desolve(F, y, [a, b])`, which involves the differential equation that we want to solve (this is denoted by  $F$  and it will necessarily include the first derivative `diff(y, x)` of  $y = f(x)$ ), the function  $y$ , which we should first introduce as a symbolic function in Sage, and the numbers  $a, b$  which are specified by the initial condition  $y(a) = b$ . For our case this technique takes the form

```
y = function("y")(x)
difeq = diff(y, x) - (2/5)*x == 0
h = desolve(difeq, y, [1, 3])
show(h)
```

Sage's output has the desired form, i.e.,  $\frac{1}{5}x^2 + \frac{14}{5}$ . □

A very elementary method of integration is the so called *integration by parts* (see 6.2.4). This method is appropriate for computing integrals of the following form:

$$\begin{aligned} & \int P(x) a^{bx} dx, & \int P(x) \sin(bx) dx, \\ & \int P(x) \cos(bx) dx, & \int P(x) \log_a^n x dx, \\ & \int x^b \log_a^n(kx) dx, & \int P(x) \arcsin(bx) dx, \\ & \int P(x) \arccos(bx) dx, & \int P(x) \arctan(bx) dx, \\ & \int P(x) \operatorname{arccot}(bx) dx, & \int a^{bx} \sin(cx) dx, \end{aligned}$$

and  $\int a^{bx} \cos(cx) dx$ , where  $P$  is an arbitrary polynomial,  $a \in (0, 1) \cup (1, +\infty)$ ,  $b, c \in \mathbb{R} \setminus \{0\}$ ,  $n \in \mathbb{N}$  and  $k > 0$ . Let us now illustrate this method by a series of examples.

**6.B.6. Integration by parts.** Using integration by parts, evaluate the integral  $L = \int (x^2 + 1) e^{-x} dx$ , with  $x \in \mathbb{R}$ .

**Solution.** Here we should apply integration by parts twice:

$$\begin{aligned} L &= \int (x^2 + 1) (-e^{-x})' dx = -(x^2 + 1) e^{-x} \\ &\quad - \int (x^2 + 1)' (-e^{-x}) = -(x^2 + 1) e^{-x} + 2\ell, \end{aligned}$$

where  $\ell = \int x e^{-x} dx$ . Next we have

$$\begin{aligned} \ell &= \int x(-e^{-x})' = -x e^{-x} - \int (x)'(-e^{-x}) \\ &= -x e^{-x} + \int e^{-x} = -x e^{-x} - e^{-x} + C, \end{aligned}$$

for some constant  $C \in \mathbb{R}$ . Thus, all together we obtain  $L = -e^{-x}(x^2 + 2x + 3) + C$ . □

**6.B.7.** Using integration by parts, compute the indefinite integrals given below:

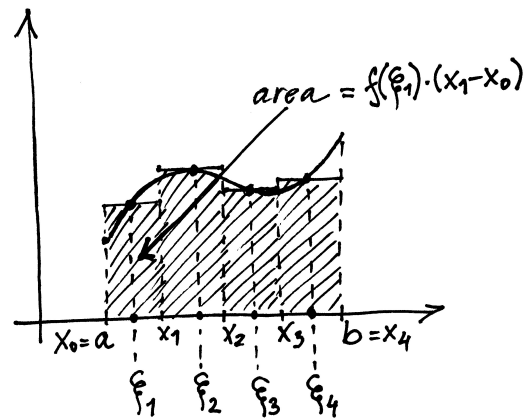
(a)  $K = \int x \cos(x) dx$ , with  $x \in \mathbb{R}$ ;

neighbourhood these functions are unbounded. We return to this problem later (see paragraph 6.2.14 below).

**6.2.8. Riemann integral.** We return to the idea of defining the integral as a tool for computing the area of the region bounded by the graph of a function and the  $x$  axis. This is our next goal. We prove that for all continuous functions on a closed bounded interval, this definition yields the same result as the Newton integral.

Consider a real function  $f$  defined on the interval  $[a, b]$ . Choose a partition of this interval along with the choice of representatives  $\xi_i$  of the respective parts, i.e.  $a = x_0 < x_1 < \dots < x_n = b$  and  $\xi_i \in [x_{i-1}, x_i]$ ,  $i = 1, \dots, n$ . The number  $\delta = \min_i \{x_i - x_{i-1}\}$  is called the *norm of the partition*. Define the *Riemann sum* corresponding to the chosen partition along with the chosen representatives  $\Xi = (x_0, \dots, x_n; \xi_1, \dots, \xi_n)$  as

$$S_{\Xi} = \sum_{i=1}^n f(\xi_i) \cdot (x_i - x_{i-1}).$$



RIEMANN INTEGRAL<sup>4</sup>

**Definition.** The *Riemann integral* of the function  $f$  on the interval  $[a, b]$  exists, if for every sequence of partitions with representatives  $(\Xi_k)_{k=0}^{\infty}$  with norms of the partitions  $\delta_k$  approaching zero, the limit

$$\lim_{k \rightarrow \infty} S_{\Xi_k} = S$$

exists and its value does not depend on the choice of the sequence of partitions and their representatives. Then we write

$$S = \int_a^b f(x) dx.$$

This definition does not look very practical, but nonetheless it allows us to formulate and prove several simple properties of the Riemann integral:

<sup>4</sup>Bernhard Riemann (1826-1866) was an extremely influential German mathematician with many contributions to infinitesimal analysis, differential geometry, and in particular complex analysis and analytic number theory.

$$(b) M = \int (2x - 1) \ln(x) dx, \text{ with } x > 0;$$

$$(c) N = \int e^x \sin(\beta x) dx, \text{ with } x, \beta \in \mathbb{R}. \quad \circ$$

**6.B.8.** Determine the integrals given below and next use Sage to confirm your answers:

$$(a) \int \frac{x}{\cos^2(x)} dx, \text{ with } x \neq \frac{\pi}{2} + k\pi, k \in \mathbb{Z};$$

$$(b) \int x^2 e^{-3x} dx, \text{ with } x \in \mathbb{R};$$

$$(c) \int \cos^2(x) dx, \text{ with } x \in \mathbb{R}. \quad \circ$$



The substitution method, introduced in 6.2.5, is another very important technique to compute integrals. The next series of exercises is based on this method. Notice later you will meet a more systematic illustration of the substitution method, along the integration of rational and irrational functions, but also in many trigonometric integrals.

**6.B.9. Substitution method.** Using a suitable substitution, determine the integral  $\int f(x) dx$ , where  $f(x)$  is given below.

Moreover, use Sage to confirm your answer.

$$(1) f(x) = \sqrt{2x - 5}, \text{ with } x > \frac{5}{2};$$

$$(2) f(x) = (7 + \ln(x))^7 / x, \text{ with } x > 0;$$

$$(3) f(x) = \cos(x) / (1 + \sin(x))^2, \text{ with } x \neq \frac{(3+4k)\pi}{2}, k \in \mathbb{Z};$$

$$(4) f(x) = (1 - 2x)^{2025}, \text{ with } x < 1/2;$$

**Solution.** (1) Set  $t = 2x - 5$  such that  $dt = 2 dx$ , i.e.,  $dx = \frac{dt}{2}$ . Thus

$$\int \sqrt{2x - 5} dx = \frac{1}{2} \int \sqrt{t} dt = \frac{t^{\frac{3}{2}}}{\frac{3}{2}} + C = \frac{1}{3}(2x - 5)^{\frac{3}{2}} + C$$

for some constant  $C$ .

(2) Set  $t = 7 + \ln(x)$  such that  $dt = \frac{dx}{x}$ . Then we get

$$\int \frac{(7 + \ln(x))^7}{x} dx = \int t^7 dt = \frac{t^8}{8} + C = \frac{(7 + \ln(x))^8}{8} + C.$$

(3) Set  $t = 1 + \sin(x)$  such that  $dt = \cos(x) dx$ . Then

$$\int \frac{\cos(x)}{(1 + \sin(x))^2} dx = \int \frac{dt}{t^2} = -\frac{1}{t} + C = -\frac{1}{(1 + \sin(x))} + C.$$

(4) Set  $t = 1 - 2x$  such that  $dt = -2 dx$ , i.e.,  $dx = -\frac{dt}{2}$ . Then

$$\begin{aligned} \int (1 - 2x)^{2025} dx &= -\frac{1}{2} \int t^{2025} dt = -\frac{1}{2} \cdot \frac{t^{2026}}{2026} + C \\ &= -\frac{(1 - 2x)^{2026}}{4052} + C, \quad C \in \mathbb{R}. \end{aligned}$$

A verification via Sage occurs with the same method presented in 6.B.1. Hence one can type

```
show(integral(sqrt(2*x-5), x))
show(integral(((7+ln(x))^7)/x, x))
show(integral((cos(x)/(1+sin(x))^2), x))
```

**Theorem.** (1) Suppose  $f$  is a bounded real function defined on the interval  $[a, b]$ , and  $c \in [a, b]$  is an inner point of this interval. Then the integral  $\int_a^b f(x) dx$  exists if and only if both of the integrals  $\int_a^c f(x) dx$  and  $\int_c^b f(x) dx$  exist. In that case

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx.$$

(2) Suppose  $f$  and  $g$  are two real functions defined on the interval  $[a, b]$ , and that both of the integrals  $\int_a^b f(x) dx$  and  $\int_a^b g(x) dx$  exist. Then the integral of their sum also exists and

$$\int_a^b (f(x) + g(x)) dx = \int_a^b f(x) dx + \int_a^b g(x) dx.$$

(3) Suppose  $f$  is a real function defined on the interval  $[a, b]$ ,  $C \in \mathbb{R}$  is a constant, and the integral  $\int_a^b f(x) dx$  exists. Then the integral  $\int_a^b Cf(x) dx$  also exists and

$$\int_a^b Cf(x) dx = C \int_a^b f(x) dx.$$

**PROOF.** (1) First suppose that the integral over the whole interval exists. When computing it, we can limit ourselves to limits of the Riemann sums whose partitions have the point  $c$  among their partitioning points. Each such sum can be obtained as a sum of two partial Riemann sums. If these two partial sums would depend on the chosen partitions and representatives in the limit, then the total sums could not be independent on the choices in limit. (It suffices to keep the sequence of partitions of the subinterval the same, and change the other so that the limit would change).

Conversely, if both Riemann integrals on both subintervals exists, they can be approximated with arbitrary precision by the Riemann sums, and moreover independently on their choice. If a partitioning point  $c$  is added to any sequence of Riemann sums over the whole interval  $[a, b]$ , the value of the whole sum is changed. Also the values of the partial sums over the intervals belonging to  $[a, c]$  and  $[c, b]$  change at most by a multiple of the norm of the partition and possible differences of the bounded function  $f$  on all of  $[a, b]$ . This is a number arbitrarily close to zero for a decreasing norm of the partition. Necessarily the partial Riemann sums of the function over the two parts of the interval also converge to the limits, whose sum is the Riemann integral over  $[a, b]$ .

(2) In every Riemann sum, the sum of the functions manifests as the sum of the values in the chosen representatives. Because multiplication of real numbers is distributive, each Riemann sum becomes the sum of the two Riemann sums with the same representatives for the two functions. The statement follows from the elementary properties of limits.

(3) Each of the Riemann sum is multiplied by the constant  $C$ . So the claim follows from the elementary properties of limits.  $\square$

and similarly for final case. □

**6.B.10.** Using a suitable substitution determine  $\int e^{x^3+4} x^2$ , with  $x \in \mathbb{R}$ . Next use Sage to confirm the answer. ○

**6.B.11.** Compute  $\int \cos^5(x) \sin(x) dx$ , with  $x \in \mathbb{R}$ . ○

**6.B.12.** Compute  $\int \frac{\sin^4(x)}{\cos^4(x)} dx$ , with  $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$ . ○

**6.B.13.** Use Sage to confirm the answers presented in **6.B.11** and **6.B.12**, respectively. ○

**6.B.14.** Evaluate  $\int \cos^5(x) \sin^2(x) dx$ . ○

*Combining integration by parts with the substitution method is a very common situation. Hence we will meet often examples which rely on combining both methods, as the task presented below.*

**6.B.15.** Evaluate the following integrals:

(a)  $\int x^3 e^{-x^2} dx$ , with  $x \in \mathbb{R}$ ;

(b)  $\int x \arcsin(x^2) dx$ , with  $x \in (-1, 1)$ ;

(c)  $\int e^{\sqrt{x}} dx$ , with  $x > 0$ . ○

**6.B.16. Integration by reduction to recurrences.** Let  $n$  be a non-negative integer.

(a) Set  $I_n = \int x^n e^x dx$ , with  $I_0 = e^x$ . Show that  $I_n = x^n e^x - nI_{n-1}$  and next compute  $I_1, I_2, I_3$ .

(b) Set  $J_n = \int (\ln(x))^n dx$ . Show that  $J_n = x(\ln(x))^n - nJ_{n-1}$  and next compute  $J_1, J_2, J_3$ .

**Solution.** (a) Integration by parts gives the result:

$$I_n = \int x^n (e^x)' dx = x^n e^x - n \int x^{n-1} e^x dx = x^n e^x - nI_{n-1}.$$

It is easy now to compute  $I_1, I_2, I_3$ :

$$I_1 = x e^x - e^x = e^x(x - 1),$$

$$I_2 = x^2 e^x - 2e^x(x - 1) = e^x(x^2 - 2x + 2),$$

$$I_3 = x^3 e^x - 3e^x(x^2 - 2x + 2) = e^x(x^3 - 3x^2 + 6x - 6).$$

You can also prove that  $I_n = e^x p_n(x)$  where the polynomial  $p_n$  is of order  $n$  and defined recursively by

$$p_n(x) = x^n - n p_{n-1}(x), \quad p_0(x) = 1.$$

(b) This can be solved similarly and left for practice. □

**6.B.17.** Let  $I_n = \int \sin^n(x) dx$  with  $n \in \mathbb{N}$ . Prove that

$$I_n = -\frac{1}{n} \sin^{n-1}(x) \cos(x) + \frac{n-1}{n} I_{n-2}.$$

Then use this relation to compute  $I_2$  and  $I_3$ . ○

**6.2.9. The fundamental theorem.** The following result is crucial for understanding the relation between the integral and the derivative. The complete proof of this theorem is somewhat longer, so it is broken into several subsections.



FUNDAMENTAL THEOREM OF INTEGRAL CALCULUS

**Theorem.** For every continuous function  $f$  on a finite interval  $[a, b]$  there exists its Riemann integral  $\int_a^b f(x) dx$ . Moreover, the function  $F(x)$  given on the interval  $[a, b]$  by the Riemann integral

$$F(x) = \int_a^x f(t) dt$$

is a primitive function to  $f$  on this interval.

**6.2.10. Upper and lower Riemann integral.** In the first step for proving the existence of the integral, we use an alternative definition, in which the choice of representatives and the corresponding value  $f(\xi_i)$  is replaced by the suprema  $M_i$  of the values  $f(x)$  in the corresponding subintervals  $[x_{i-1}, x_i]$ , or by the infima  $m_i$  of the function  $f(x)$  in the same subintervals, respectively. We speak of upper and lower Riemann sums, respectively (in literature, this process is also called the *Darboux integral*).

Because the function is continuous, it is bounded on a closed interval, hence all the above considered suprema and infima exist and are finite. Then the *upper Riemann sum* corresponding to the partition  $\Xi = (x_0, \dots, x_n)$  is given by the expression

$$S_{\Xi, \text{sup}} = \sum_{i=1}^n \left( \sup_{x_{i-1} \leq \xi \leq x_i} f(\xi) \right) (x_i - x_{i-1}) = \sum_{i=1}^n M_i (x_i - x_{i-1}).$$

The *lower Riemann sum* is

$$S_{\Xi, \text{inf}} = \sum_{i=1}^n \left( \inf_{x_{i-1} \leq \xi \leq x_i} f(\xi) \right) (x_i - x_{i-1}) = \sum_{i=1}^n m_i (x_i - x_{i-1}).$$

For each partition  $\Xi = (x_0, \dots, x_n; \xi_1, \dots, \xi_n)$  with representatives, there are the inequalities

$$(1) \quad S_{\Xi, \text{inf}} \leq S_{\Xi, \xi} \leq S_{\Xi, \text{sup}}$$

Moreover, the infima and suprema can be approximated with arbitrary precision by the actual values of terms in the sequences. Thus, we might suspect that the Riemann integral exists if and only if for all sequences of partitions with norms approaching zero, the limits of both the upper and lower sums will exist and they will be equal. This is indeed true for all bounded functions: ○



An important situation in integration occurs along the integration of rational functions, see 6.2.7. The key point here is the decomposition of a such a function as a sum of simple rational functions. Integrating the partial fractions corresponding to real roots of a denominator of a rational function is very easy. For instance,

$$\int \frac{A}{x-x_0} dx \stackrel{y=x-x_0}{\underset{dy=x}{=}} A \ln|y| + C = A \ln|x-x_0| + C_1,$$

$$\int \frac{A}{(x-x_0)^n} dx \stackrel{y=x-x_0}{\underset{dy=x}{=}} \int \frac{A}{y^n} dy = \frac{Ay^{-n+1}}{-n+1} + C_2$$

$$= \frac{A}{(1-n)(x-x_0)^{n-1}} + C_2,$$

for some constants  $C_1, C_2$ , and for all  $A, x_0 \in \mathbb{R}, n \in \mathbb{N}$  with  $n \geq 2$ . Rational functions  $f(x)/g(x)$  can be integrated easily, assuming that we know the corresponding factorization of the polynomial in the denominator  $g(x)$  in terms of its roots and their multiplicities. Next we will cover all the possible cases and to begin with, here are a few easy tasks for you.

**6.B.18.** Evaluate the integral  $\int R(x) dx$ , for  $R(x) = \frac{6}{x-2}$ , with  $x \neq 2$  and  $R(x) = \frac{6}{(x+4)^3}$ , with  $x \neq -4$ . ○

**6.B.19.** (a) Prove that

$$\int \frac{dx}{(ax+b)^n} = \frac{1}{a(1-n)(ax+b)^{n-1}} + C.$$

(b) Compute  $\int \frac{dx}{9x^2+6x+1}$  by applying the formula in (a). ○

**6.B.20.** Suppose that  $P(x) = ax^2 + bx + c$  ( $a \neq 0$ ) is a parabola with a double root, i.e.,  $\Delta = b^2 - 4ac = 0$ . Prove that

$$\int \frac{dx}{ax^2 + bx + c} = -\frac{2}{2ax + b} + C.$$

Use this formula to confirm your result in 6.B.19, (b). ○

**6.B.21.** Evaluate the integral  $\int \frac{1}{5x^2 + 5x + 2} dx$ . Next use Sage to confirm your result.

**Solution.** The given parabola  $P(x) = x^2 + 5x + 2$  has negative discriminant  $\Delta = -15$ , and hence two complex conjugate roots, given by  $-\frac{1}{2} \pm i\frac{\sqrt{15}}{10}$ . By completing the square we get

$$P(x) = 5\left(x^2 + x + \frac{2}{5}\right) = 5\left(\left(x + \frac{1}{4}\right) - \frac{1}{4} + \frac{2}{5}\right)$$

$$= 5\left(\left(x + \frac{1}{2}\right)^2 + \frac{3}{20}\right) = 5\left(\left(x + \frac{1}{2}\right)^2 + \left(\sqrt{3/20}\right)^2\right)$$

$$= 5\left(\left(x + \frac{1}{2}\right)^2 + \left(\sqrt{15/10}\right)^2\right).$$

**Theorem.** Let the function  $f$  be bounded on a closed interval  $[a, b]$ . Then

$$S_{\text{sup}} = \inf_{\Xi} S_{\Xi, \text{sup}}, \quad S_{\text{inf}} = \sup_{\Xi} S_{\Xi, \text{inf}}$$

are the limits of all sequences of upper and lower sums with norm approaching zero, respectively.

The Riemann integral of the function  $f$  exists if and only if  $S_{\text{sup}} = S_{\text{inf}}$ .

**PROOF.** First, notice that choosing two partitions  $\Xi_1, \Xi_2$ , there is the common refinement  $\Xi$  and we arrive at the inequalities  $S_{\Xi_1, \text{inf}} \leq S_{\Xi, \text{inf}} \leq S_{\Xi, \text{sup}} \leq S_{\Xi_2, \text{sup}}$ . Thus,  $S_{\text{sup}}$  is well defined since it is the infimum of a set of real values bounded from below by any of the  $S_{\Xi, \text{inf}}$ . Similarly for the value  $S_{\text{inf}}$ , which is bounded from above by any of  $S_{\Xi, \text{sup}}$ .

Refine a partition  $\Xi_1$  to  $\Xi_2$  by adding new points. Then

$$S_{\Xi_1, \text{sup}} \geq S_{\Xi_2, \text{sup}}, \quad S_{\Xi_1, \text{inf}} \leq S_{\Xi_2, \text{inf}}.$$

By the definition of the infimum, there are sequences of partitions  $\Xi_k$  for which  $S_{\text{sup}}$  is the limit of the sums  $S_{\Xi_k, \text{sup}}$ . Moreover, every two partitions have a common refinement. Thus it may be assumed that  $\Xi_k$  in the sequence is always a partition obtained by refining the previous one. Hence the sums  $S_{\Xi_k, \text{sup}}$  form a non-increasing sequence of real numbers converging to  $S_{\text{sup}}$ .

A similar argument applies to  $S_{\text{inf}}$ . Hence the values

$$S_{\text{sup}} = \inf_{\Xi} S_{\Xi, \text{sup}}, \quad S_{\text{inf}} = \sup_{\Xi} S_{\Xi, \text{inf}}$$

are good candidates for the limits of upper and lower sums.

Next, consider a fixed partition  $\Xi$  with  $n$  inner partitioning points of the interval  $[a, b]$ , and another partition  $\Xi_1$ , whose norm is a small number  $\delta$ . In the common refinement  $\Xi_2$ , there will be only  $n$  intervals contributing to the sum  $S_{\Xi_2, \text{sup}}$  by eventually smaller contribution than in the case of  $\Xi_1$ . Now,  $f$  is a bounded function on  $[a, b]$  and thus each of these contributions will be bounded by a universal constant multiplied by the norm  $\delta$  of the partition. Hence when choosing  $\delta$  sufficiently small, the distance of  $S_{\Xi_1, \text{sup}}$  from  $S_{\text{sup}}$  will not be larger than twice the distance of  $S_{\Xi, \text{sup}}$  from  $S_{\text{sup}}$ .

Finally, return to the sequence of partitions  $\Xi_k$  as chosen above, and choose an  $\varepsilon > 0$ . Then there is some  $m \in \mathbb{N}$  such that the distance of  $S_{\Xi_k, \text{sup}}$  from  $S_{\text{sup}}$  is less than  $\varepsilon$  for all  $k \geq m$ . Hence for arbitrary partition  $\Xi$  with appropriately small norm  $\delta > 0$  the distance of  $S_{\Xi, \text{sup}}$  from  $S_{\text{sup}}$  does not exceed  $2\varepsilon$ .

In summary, for arbitrary  $2\varepsilon > 0$ , there is  $\delta > 0$  such that for all partitions with norm at most  $\delta$  the inequality  $|S_{\Xi, \text{sup}} - S_{\text{sup}}| < 2\varepsilon$  holds. This is exactly the statement that the number  $S_{\text{sup}}$  is the limit of all sequences of upper sums with norms of the partition approaching zero.

The statement for lower sums is proved in exactly the same way.

It remains to deal with the existence of the Riemann integral  $\int_a^b f(x) dx$ . If  $S_{\text{sup}} = S_{\text{inf}}$ , then all Riemann sums of

Thus  $\int \frac{1}{P(x)} dx = \frac{1}{5} \int \frac{dx}{(x + \frac{1}{2})^2 + (\sqrt{15}/10)^2}$ . Next, set  $u = x + \frac{1}{2}$  with  $du = dx$ , and use the known formula (see 6.2.3)

$$\int \frac{dx}{x^2 + \delta^2} = \frac{1}{\delta} \arctan\left(\frac{u}{\delta}\right) + C. \quad (\#)$$

This gives

$$\begin{aligned} \frac{1}{5} \int \frac{dx}{(x + \frac{1}{2})^2 + (\sqrt{15}/10)^2} &= \frac{1}{5} \int \frac{du}{u^2 + (\sqrt{15}/10)^2} \\ &= \frac{1}{5} \cdot \frac{1}{\sqrt{15}/10} \arctan\left(\frac{u}{\sqrt{15}/10}\right) + C \\ &= \frac{2\sqrt{15}}{15} \arctan\left((\sqrt{15}/3)(2x + 1)\right) + C, \end{aligned}$$

for some constant  $C$ . For a confirmation give the cell show(integral(1/(5 \* x^2 + 5 \* x + 2), x)).  $\square$

*Integrals of a rational function  $Q$  of the form  $Q(x) = (a_1x + b_1)/(a_2x^2 + b_2x + c_2)$  can be solved by forming on the numerator the derivative of the dominator. Then we decompose the fraction into two rational functions and hence our integral decomposes into two integrals, that are easier to be computed. Let us describe such an example.*

**6.B.22.** Evaluate the integral  $A = \int \frac{3x + 7}{x^2 - 4x + 15} dx$ , with  $x \in \mathbb{R}$ .

**Solution.** The derivative of  $g(x) = x^2 - 4x + 15$  is the line  $2x - 4$ , with  $x \in \mathbb{R}$ . Let  $f(x) = 3x + 7$ , with  $x \in \mathbb{R}$ , as well. We see that  $f(x) = \frac{3}{2}(2x - 4) + 4 \cdot \frac{3}{2} + 7 = \frac{3}{2}(2x - 4) + 13$ . Thus

$$A = \int \frac{f(x)}{g(x)} dx = \frac{3}{2} \int \frac{2x - 4}{g(x)} dx + 13 \int \frac{dx}{g(x)} = \alpha + \beta.$$

For  $\alpha := \frac{3}{2} \int \frac{2x - 4}{g(x)} dx$ , by setting  $u = g(x) = x^2 - 4x + 15$  we get  $du = (2x - 4) dx$  and thus

$$\alpha = \frac{3}{2} \int \frac{du}{u} = \frac{3}{2} \ln(x^2 - 4x + 15) + C_1, \quad C_1 \in \mathbb{R}.$$

For the second integral  $\beta := 13 \int \frac{dx}{g(x)}$  by completing the square we get  $g(x) = (x - 2)^2 + 11$ . Thus an application of the relation (#) appearing in the proof of 6.B.21, gives

$$\beta = 13 \int \frac{dx}{(x - 2)^2 + (\sqrt{11})^2} = \frac{13}{\sqrt{11}} \arctan\left(\frac{x - 2}{\sqrt{11}}\right) + C_2$$

for some constant  $C_2$ . Hence all together we get that  $A = \frac{3}{2} \ln(x^2 - 4x + 15) + \frac{13}{\sqrt{11}} \arctan\left(\frac{x - 2}{\sqrt{11}}\right) + C$ .  $\square$

**6.B.23.** If  $K_n = K_n(x_0, a) = \int \frac{dx}{((x - x_0)^2 + a^2)^n}$ , where  $a, x_0 \in \mathbb{R}$  are fixed, prove that

$$K_n = \frac{1}{a^2} \left( \frac{2n - 3}{2n - 2} K_{n-1} + \frac{x - x_0}{(2n - 2)((x - x_0)^2 + a^2)^{n-1}} \right)$$

sequences of the partitions have the same limit because of the inequalities (1).

If the Riemann integral does not exist, then there exist two sequences of partitions  $\Xi_k$  and  $\bar{\Xi}_k$  and their representatives with different limits of Riemann sums. Suppose the first limit is larger then the other one. Then the upper Riemann sums can be selected for the first sequence and the lower Riemann sums for the second sequence. Their difference will then be at least as large. In particular, in view of the previous part of the proof, this implies  $S_{\text{sup}} > S_{\text{inf}}$ .  $\square$

**6.2.11. Uniform continuity.** Until now, we have only used the continuity of the function  $f$  to know that all such functions are bounded on a closed finite interval. It remains to show that for continuous functions  $S_{\text{sup}} = S_{\text{inf}}$ .



From the definition of continuity, for every point  $x \in [a, b]$  and every neighbourhood  $\mathcal{O}_\varepsilon(f(x))$  there exists a neighbourhood  $\mathcal{O}_\delta(x)$  such that  $f(\mathcal{O}_\delta(x)) \subset \mathcal{O}_\varepsilon(f(x))$ . This statement can be rewritten in this way: for  $y, z \in \mathcal{O}_\delta(x)$ , i.e.

$$|y - z| < 2\delta,$$

it is true that  $f(y), f(z) \in \mathcal{O}_\varepsilon(f(x))$ , i.e.

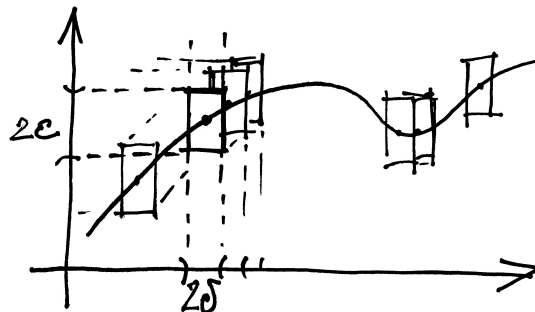
$$|f(y) - f(z)| < 2\varepsilon.$$

A global variant of such a property is needed; it is called the *uniform continuity* of a function  $f$ :

#### UNIFORM CONTINUITY

**Definition.** Let  $f$  be a function on a closed finite interval  $[a, b]$ .  $f$  is *uniformly continuous* on  $[a, b]$ , if for every  $\varepsilon > 0$  there exists  $\delta > 0$  such that for all  $z, y \in [a, b]$  satisfying  $|y - z| < \delta$ , the inequality  $|f(y) - f(z)| < \varepsilon$  holds.

#### UNIFORM CONTINUITY



**Theorem.** Each continuous function on a finite closed interval  $[a, b]$  is *uniformly continuous*.

**PROOF.** Fixing some  $\varepsilon > 0$ , the definition of continuity of  $f$  provides for each  $x \in [a, b]$  the values  $\delta(x)$ , such that  $f(y) \in \mathcal{O}_\varepsilon(f(x))$  for all  $y \in \mathcal{O}_{2\delta(x)}(x)$ . Since every finite closed interval is compact, it is covered by finitely many of such neighbourhoods  $\mathcal{O}_{\delta(x)}(x)$ , determined by points

for all  $n \neq 1$ , and that  $K_1(x_0, a) = \frac{1}{a} \arctan\left(\frac{x-x_0}{a}\right) + C$ .

**Solution.** We will apply integration by parts. In terms of 6.2.4 we have  $F(x) = 1/((x-x_0)^2 + a^2)^n$  and  $F'(x) = -2n(x-x_0)/((x-x_0)^2 + a^2)^{n+1}$ . If you are not sure for this differentiation, use Sage and the cell

```
var("a, n, x0"); F(x)=1/((x-x0)^2+a^2)**n
show(diff(F(x), x).factor())
```

Sage prints out the following expression which agrees with ours:  $-2(a^2 + x^2 - 2xx_0 + x_0^2)^{-n-1}n(x-x_0)$ . Moreover,  $G'(x) = 1$  and we may fix  $G(x) = (x-x_0)$ . Thus,

$$K_n(x_0, a) = \int F(x)G'(x) dx = \frac{x-x_0}{((x-x_0)^2 + a^2)^n} + 2n \int \left( \frac{(x-x_0)^2 + a^2}{((x-x_0)^2 + a^2)^{n+1}} - \frac{a^2}{((x-x_0)^2 + a^2)^{n+1}} \right) dx$$

$$= \frac{x-x_0}{((x-x_0)^2 + a^2)^n} + 2n(K_n(x_0, a) - a^2 K_{n+1}(x_0, a)),$$

or equivalently,

$$K_{n+1} = \frac{1}{a^2} \left( \frac{2n-1}{2n} K_n + \frac{1}{2n} \frac{x-x_0}{((x-x_0)^2 + a^2)^n} \right).$$

Replacing  $n$  by  $n-1$  in this formula, we get the result. On the other hand, the case  $n=1$  follows easily by the relation (#), used in the proof of 6.B.21.  $\square$

**6.B.24.** Using the result from 6.B.23, compute the integral given below and next verify your answer via Sage:

$$I = \int \frac{30x - 77}{(x^2 - 6x + 13)^2} dx, \quad x \in \mathbb{R}.$$

**Solution.** This provides an example of partial fractions for multiple complex roots in the form of

$$\frac{Ax+B}{[(x-x_0)^2 + a^2]^n}, \quad \text{with } A, B, x_0 \in \mathbb{R}, a > 0, n \in \mathbb{N} \setminus \{0, 1\},$$

which again can be solved by forming on the numerator the derivative of the expression  $(x-x_0)^2 + a^2$  appearing in the denominator, that is,  $\frac{A}{2} \cdot \frac{2(x-x_0)}{[(x-x_0)^2 + a^2]^n} + (B + Ax_0) \cdot \frac{1}{[(x-x_0)^2 + a^2]^n}$ . Hence one needs to compute the two induced integrals, which can be done by the methods applied above. For our problem we have

$$I = 15 \int \frac{2x-6}{(x^2-6x+13)^2} dx + 13 \int \frac{dx}{(x^2-6x+13)^2}.$$

To compute the first integral set  $u = x^2 - 6x + 13$  with  $du = (2x-6) dx$ . Then

$$15 \int \frac{2x-6}{(x^2-6x+13)^2} dx = 15 \int \frac{du}{u^2} = -\frac{15}{u} + C_1 = -\frac{15}{x^2-6x+13} + C_1,$$

for some constant  $C_1$ . For the second integral by completing the square one gets  $x^2 - 6x + 13 = (x-3)^2 + 2^2$ . Hence

$$13 \int \frac{1}{(x^2-6x+13)^2} dx = 13 \int \frac{dx}{((x-3)^2 + 2^2)^2}.$$

$x_1, \dots, x_k$ . Choose  $\delta$  as the minimum of all the (finitely many)  $\delta(x_i)$ .

Choose any two points  $y, z \in [a, b]$  with  $|y-z| < \delta$ , they both belong to one of  $\mathcal{O}_{2\delta(x_i)}(x_i)$ . Thus  $|f(y)-f(z)| \leq |f(y)-f(x_i)| + |f(x_i)-f(z)| < 2\varepsilon$  and  $f$  has the desired property.  $\square$

**6.2.12. Finishing the proof of Theorem 6.2.9.** Now we complete the proof of the existence of the Riemann integral. Choose  $\varepsilon$  and  $\delta$  as in the definition of the uniform continuity of  $f$ . Consider any partition  $\Xi$  with  $n$  intervals and norm at most  $\delta$ . Then, writing  $J_i = [x_{i-1}, x_i]$ ,



$$\left| \sum_{i=1}^n \sup_{\xi \in J_i} f(\xi)(x_i - x_{i-1}) - \sum_{i=1}^n \inf_{\xi \in J_i} f(\xi)(x_i - x_{i-1}) \right|$$

$$= \sum_{i=1}^n (\sup_{\xi \in J_i} f(\xi) - \inf_{\xi \in J_i} f(\xi))(x_i - x_{i-1})$$

$$\leq \varepsilon \cdot (b-a).$$

For decreasing norm of the partition, the upper and lower sums are arbitrarily close to each other. In particular the upper Riemann integral and the lower Riemann integral coincide.

To complete the proof of the fundamental theorem of integral calculus, it is still needed to verify the statement about the existence of a primitive function. For a continuous function  $f$  on interval  $[a, b]$  there exists the Riemann integral  $F(x) = \int_a^x f(t) dt$  for every  $x \in [a, b]$ . As in the statement about uniform continuity, there is  $\delta > 0$ , dependent on a fixed small  $\varepsilon > 0$ , such that

$$|f(x + \Delta x) - f(x)| < \varepsilon$$

for all  $0 \leq \Delta x < \delta$  on the interval  $[a, b]$ . The difference of the derivative of  $F(x)$  and the integrated function  $f(x)$  is expressed by the limit of the expressions

$$\alpha = \frac{1}{\Delta x} \left( \int_a^{x+\Delta x} f(t) dt - \int_a^x f(t) dt \right) - f(x)$$

$$= \frac{1}{\Delta x} \left( \int_x^{x+\Delta x} f(t) dt \right) - f(x)$$

for  $\Delta x$  approaching zero.

Now choose  $0 < \Delta x < \delta$  and replace the integrated function by the constant value  $f(x)$ . Then the values  $f(\xi)$  at any point  $\xi \in [x, x + \Delta x]$  are distant from  $f(x)$  by at most  $\varepsilon$ . Hence the Riemann integral in question cannot be different from  $f(x)\Delta x$  by more than  $\varepsilon\Delta x$ . Thus, we arrived at the following estimate:

$$|\alpha| = \left| \frac{1}{\Delta x} \left( \int_x^{x+\Delta x} f(t) dt \right) - f(x) \right| < \varepsilon.$$

But that means that at the point  $x$ , the one-sided right derivative of the function  $F(x)$  exists and equals  $f(x)$ . The result for the left derivative is proved in the same way, just working with the interval  $[x - \Delta x, x]$ . This finishes the proof of the theorem 6.2.9.

We can now apply 6.B.23 with  $x_0 = 3$ ,  $a = 2$  and  $n = 2$ . In particular we have

$$\begin{aligned} 13 \cdot K_2(3, 2) &= 13 \int \frac{dx}{((x-3)^2 + 2^2)^2} \\ &= \frac{13}{2^2} \left( \frac{1}{2} K_1(3, 2) + \frac{x-3}{2((x-3)^2 + 2^2)} \right) \\ &= \frac{13}{4} \left( \frac{1}{4} \arctan \left( \frac{x-3}{2} \right) + C_2 + \frac{x-3}{2((x-3)^2 + 2^2)} \right) \\ &= \frac{13}{16} \arctan \left( \frac{x-3}{2} \right) + \frac{13(x-3)}{8(x^2 - 6x + 13)} + C_2, \end{aligned}$$

for some constant  $C_2$ . In total we get

$$I = \frac{13}{16} \arctan \left( \frac{x-3}{2} \right) + \frac{13x-159}{8(x^2 - 6x + 13)} + C, \quad C \in \mathbb{R}.$$

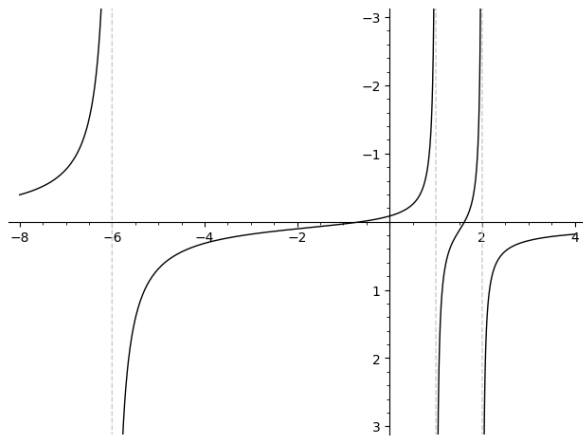
A verification via Sage occurs as usual, i.e., type `show(integral((30 * x - 77)/(x^2 - 6 * x + 13)^2, x))`. □

**6.B.25.** Consider the rational function

$$R(x) = \frac{x^2 - x - 1}{x^3 + 3x^2 - 16x + 12}.$$

- (a) Find its domain and its discontinuities and sketch its graph.
- (b) Find the decomposition of  $R$  into partial fractions.
- (c) Evaluate the integral  $\int R(x) dx$ .

**Solution.** (a) Let us express  $R$  as  $R(x) = f(x)/g(x)$ . It is obvious that  $g(1) = 0$  and based on the Horner's scheme we get the factorization  $g(x) = (x-1)(x-2)(x+6)$ . Using Sage one can type `factor(x**3 + 3*x**2 - 16*x + 12)`. Hence, the domain of  $R$  is the set  $\mathbb{R} \setminus \{-6, 1, 2\}$ . Notice the dominator  $g$  has three distinct real roots and none of them is a root of the numerator  $f$ . Hence, all the discontinuities of  $R = f/g$  appear at the roots 1, 2 and  $-6$  of  $g$ , as we can see also from the graph of  $R$ :



(b) Let us assume that

$$R(x) = \frac{x^2 - x - 1}{(x-1)(x-2)(x+6)} = \frac{A_1}{x-1} + \frac{A_2}{x-2} + \frac{A_3}{x+6} \quad \int_a^b f(x) dx = F(b) - F(a) = F'(c)(b-a) = f(c)(b-a).$$

**6.2.13. Important remarks.** (1) Theorems 6.2.9 and 6.2.8 claim that the Riemann integral is a linear map



$$\int : C[a, b] \rightarrow \mathbb{R}$$

from the vector space of continuous functions on the interval  $[a, b]$  to real numbers. Hence it is a linear form on the vector space  $C[a, b]$ .

(2) We proved that every continuous function is a derivative of some function. Hence the concepts of the Newton and Riemann integrals coincide for continuous functions. Therefore the Riemann integral of continuous functions can be computed as the difference of values  $F(b) - F(a)$  of the primitive function  $F$ .

(3) In the first step of the proof of the theorem 6.2.9 we proved the important statement that for bounded functions  $f$  on finite intervals  $[a, b]$  the limits of the upper and lower sums always exist. They are called respectively *the upper Riemann integral* and *the lower Riemann integral* and they are also denoted by  $\overline{\int}_a^b f(x) dx$  and  $\underline{\int}_a^b f(x) dx$ .

In this way we can define the Riemann integral for continuous functions as in the above proof.

(4) We derived the important property of continuous functions called *the uniform continuity* on finite closed intervals  $[a, b]$ . Clearly every uniformly continuous function is continuous as well, but the converse is not true on open intervals. As an example, consider the function  $f(x) = \sin(1/x)$  on the interval  $(0, 1)$ .

(5) Consider a function  $f$  on an interval  $[a, b]$ , which is only *piece-wise continuous*. This means that  $f$  is continuous in all points  $c \in [a, b]$  except for finitely many *discontinuities*  $c_i$ ,  $a < c_i < b$ , in which it has finite one-sided limits. Because of the additivity of the integral with respect to the interval of integration, see 6.2.8(1), the last theorem implies that in this case the integral

$$F(x) = \int_a^x f(t) dt$$

exists for all  $x \in [a, b]$  and the derivative of the function  $F(x)$  exists at all points  $x$  where  $f$  is continuous. It can be verified that  $F(x)$  is continuous at the remaining points. So it is a continuous function on the whole interval  $[a, b]$ . When evaluating the integral by primitive functions, it is necessary to choose its individual parts so that they are connected continuously at the points  $c_i$ . Then the entire integral can be again computed as a difference of the function  $F(x)$  at its boundary values.

(6) Lagrange's mean value theorem for differentiable functions has an analogue which is called *the integral mean value theorem*. Suppose  $f(x)$  is continuous on an interval  $[a, b]$  and its primitive function is  $F(x)$ . The mean value theorem claims that there exists a point  $c$ ,  $a < c < b$  such that

for some reals  $A_1, A_2, A_3$  to be specified. This relation can be equivalently written as (by multiplying by the denominator)

$$f(x) = A_1(x-2)(x+6) + A_2(x-1)(x+6) + A_3(x-1)(x-2).$$

There are many approaches to compute  $A_1, A_2, A_3$ . For instance, we may gather together like powers of the variable and equate their coefficients. This gives a set of equations to solve for  $A_1, A_2, A_3$ . As an alternative we can plug  $x = 1$  into this equation, which gives  $A_1 = 1/7$ . Likewise, the substitution of  $x = 2$  gives  $A_2 = 1/8$  while those of  $x = -6$  yields  $A_3 = 41/56$ .

Thus, the decomposition of  $R$  into partial fractions is given by

$$R(x) = \frac{1}{7(x-1)} + \frac{1}{8(x-2)} + \frac{41}{56(x+6)}. \quad (*)$$

Sage provides a built-in method for computing the decomposition of a rational function into partial fractions, given by the command `partial_fraction`. For instance, for a Sage confirmation of (\*) one may use the cell

```
f(x)=(x^2-x-1); g(x)=(x-1)*(x-2)*(x+6)
R(x)=f(x)/g(x); show(R.partial_fraction())
```

(c) This is easy and relies on our previous conclusion (\*), i.e.,

$$\begin{aligned} \int R(x) dx &= \frac{1}{7} \int \frac{1}{x-1} dx + \frac{1}{8} \int \frac{1}{x-2} dx \\ &+ \frac{41}{56} \int \frac{1}{x+6} dx = \frac{1}{7} \ln|x-1| + \\ &+ \frac{1}{8} \ln|x-2| + \frac{41}{56} \ln|x+6| + C, \end{aligned}$$

for some constant  $C$ . Adding in the previous cell the command `show(integral(R,x))`, Sage confirms our answer.  $\square$

**6.B.26.** Evaluate the integral  $\int Q(x) dx$ , where

$$Q(x) = \frac{x}{(x-1)^2(x^2+2x+2)}, \quad x \neq 1.$$

**Solution.** According to the discussion in 6.2.7, we may assume that

$$Q(x) = \frac{A}{x-1} + \frac{B}{(x-1)^2} + \frac{Cx+D}{x^2+2x+2},$$

for  $A, B, C, D \in \mathbb{R}$ . This can be equivalently expressed as

$$x = A(x-1)(x^2+2x+2) + B(x^2+2x+2) + (Cx+D)(x-1)^2.$$

By setting  $x = 1$  we immediately get  $B = 1/5$ . By comparing the coefficients at the same powers of the polynomials we also get  $A = 1/25$ ,  $C = -1/25$ , and  $D = -8/25$ . Hence

$$Q(x) = \frac{1}{25(x-1)} + \frac{1}{5(x-1)^2} - \frac{x+8}{25(x^2+2x+2)}. \quad (*)$$

An easy way to verify this fractional decomposition is by Sage, via the function `partial_fraction` as in 6.B.25, i.e.,

This statement can be derived directly from the definition of the Riemann integral. It can be then used in the final step of the proof of the fundamental theorem of integral calculus.

**6.2.14. Improper integrals.** When discussing the integration of rational functions  $f$ , there is a need



to consider definite integrals over intervals, where  $f(x)$  has improper (one-sided) limits. Here  $f$  is neither continuous nor bounded. Thus earlier definitions and results may not apply. We speak of “improper” integrals.

A simple solution is to discuss the definite integral on a smaller sub-interval, and determine whether the limit value of such a definite integral exists when the boundary approaches the problematic point. If it does, the corresponding improper integral exists and equals this limit. We illustrate this procedure by an example:

$$I = \int_0^2 \frac{dx}{(2-x)^{1/4}}.$$

This is an improper integral, because the integrand  $f(x) = (2-x)^{-1/4} = \frac{1}{\sqrt[4]{2-x}}$  has its left-sided limit  $\infty$  at the point  $b = 2$ . The integrand is continuous at all other points. Thus, for  $0 < \delta < 2$ , consider the integrals (substituting  $y = 2-x$ )

$$\begin{aligned} I_\delta &= \int_0^{2-\delta} \frac{dx}{\sqrt[4]{2-x}} = \int_\delta^2 y^{-1/4} dy \\ &= \left[ \frac{4}{3} y^{3/4} \right]_\delta^2 = \frac{4}{3} [2^{3/4} - \delta^{3/4}]. \end{aligned}$$

Notice that  $dy = -dx$  and  $x = 2 - \delta$  corresponds to  $y = \delta$ . When  $x = 0$ , then  $y = 2$ .

The limit when  $\delta \rightarrow 0$  from the right clearly exists, so the improper integral is evaluated.

$$I = \int_0^2 \frac{dx}{\sqrt[4]{2-x}} = \frac{4}{3} 2^{3/4}.$$

We proceed in the same way to integrate over an unbounded interval. In this case, we speak of *improper Riemann integrals of the first kind*. The integrals of unbounded functions on finite intervals are *improper Riemann integrals of the second kind*.

More explicitly, we can define the integrals of both kinds as follows. For  $a \in \mathbb{R}$  and  $f$  defined on  $[a, b)$  and bounded on each  $[a, c] \subset [a, b)$ ,

$$I = \int_a^b f(x) dx = \lim_{c \rightarrow b} \int_a^c f(x) dx,$$

if the integrals and limit on the right hand side exist. Here  $b$  is either finite or  $b = \infty$ . Similarly we can have a finite fixed upper bound and  $-\infty \leq a < c < b$  and infinite lower bound. If both  $a$  and  $b$  are infinite, we can evaluate the integral as a sum of two integrals with a chosen fixed bound in the middle as in

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^a f(x) dx + \int_a^{\infty} f(x) dx.$$

```
Q(x)=x/((x-1)^2*(x^2+2*x+2))
show(R.partial_fraction())
```

Based now on (\*) we can write

$$\int Q(x) dx = \int \frac{dx}{25(x-1)} + \int \frac{dx}{5(x-1)^2} - \int \frac{x+8}{25(x^2+2x+2)} dx = \frac{1}{25} \ln|x-1| - \frac{1}{5(x-1)} - \frac{1}{50} \ln(x^2+2x+2) - \frac{7}{25} \arctan(x+1) + C,$$

for some constant  $C$ . Here, as in 6.B.22 we get

$$\int \frac{x+8}{x^2+2x+2} dx = \int \left( \frac{\frac{1}{2}(2x+2)}{x^2+2x+2} + \frac{7}{x^2+2x+2} \right) dx = \frac{1}{2} \int \frac{2x+2}{x^2+2x+2} dx + 7 \int \frac{1}{(x+1)^2+1} dx = \frac{1}{2} \ln(x^2+2x+2) + 7 \arctan(x+1) + C.$$

**6.B.27.** Evaluate the integral  $A = \int Q(x) dx$ , where

$$Q(x) = \frac{2x^4 + 2x^2 - 5x + 1}{x(x^2 - x + 1)^2}, \quad x \neq 0.$$

Next confirm your result in Sage. □ ○ ○

**6.B.28.** Evaluate  $\int \frac{1}{e^{3x} - 2e^{2x}} dx$ , with  $x \neq \ln(2)$ .

*Let us also integrate an improper rational function  $f(x)/g(x)$ , which means that the degree of the polynomial  $f$  in the numerator is greater or equal to the degree of the polynomial  $g$  in the denominator. In such a case we should first carry out the division with remainder.*

**6.B.29.** Evaluate  $\int \frac{x^3 + 2x^2 + x - 1}{x^2 - x + 1} dx$ , with  $x \in \mathbb{R}$ .

Can you confirm all the steps of your proof via Sage?

**Solution.** The degree of the numerator  $f(x) = x^3 + 2x^2 + x - 1$  is bigger from the degree of the dominator  $g(x) = x^2 - x + 1$ , hence we proceed by division of polynomials:

$$\begin{array}{r} x^3 + 2x^2 + x - 1 \\ x^2 - x + 1 \overline{) \phantom{x^3 + 2x^2 + x - 1}} \\ \underline{-x^3 + x^2 - x} \phantom{- 1} \\ 3x^2 - 1 \\ \underline{-3x^2 + 3x - 3} \\ 3x - 4 \end{array}$$

This means that  $f(x) = g(x)(x+3) + 3x-4$  and implies the following decomposition

$$\frac{f(x)}{g(x)} = (x+3) + \frac{3x-4}{x^2-x+1}. \quad (b)$$

The division of polynomials can be successfully performed in Sage, and a way goes as follows

Its existence and its value do not depend on the choice of such bound, because by changing it, we only change both summands by the same finite value, but with opposite sign.

At the same time a limit for which the upper and lower bound would approach  $\pm\infty$  at the same speed can lead to different results! For example

$$\int_{-a}^a x dx = \left[ \frac{1}{2}x^2 \right]_{-a}^a = 0,$$

even though the values of the integrals  $\int_a^b x dx$  with  $a$  fixed and  $b \rightarrow \infty$  diverge to infinity. (This is the typical behavior for all odd functions.)

Clearly, if  $f$  is continuous on  $[a, b)$ , then the Newton integral  $F(x) = \int_a^x f(y) dy$  exists for all  $x \in [a, b)$  and the improper Riemann integral exists if and only if  $\lim_{x \rightarrow b} F(x)$  exists, and its value is this limit. Thus, for continuous  $f(x)$  as in our simple example above, we compute the definite Newton integral and take its limit. This is what we did. □

The integrated functions may have more discontinuities with infinite one-sided limits and the interval of integration may be unbounded. Then the integration intervals must be split in such a way that the individual intervals of integration include only one of the above phenomena.

Hence when evaluating the improper integral of a rational function, divide the given interval according to the discontinuities of the integrated function. Then compute all the improper integrals separately. ○ ○

**6.2.15. Mean value of a function.** For a finite set of  $n$  numbers, the mean value, or arithmetic mean, is obtained by summing the numbers and dividing by  $n$ . For a Riemann integrable function  $f(x)$  on an interval (finite or infinite)  $[a, b]$ , the *mean value* is defined by

$$m(f) = \frac{1}{b-a} \int_a^b f(x) dx.$$

By definition,  $m(f)$  is the altitude of the rectangle over the interval  $[a, b]$ , which has the same area as that of the region between the  $x$  axis and the graph of the  $f(x)$  (counted with signs according to being above or below the  $x$ -axis). Hence the *integral mean value theorem* is true in general:

**Proposition.** If  $f(x)$  is a Riemann integrable function on an interval  $[a, b]$ , then there exists a number  $m(f)$  satisfying

$$\int_a^b f(x) dx = m(f)(b-a).$$

**6.2.16. Integral criterion for series.** Using the improper integral, we can also decide the question of convergence for a class of infinite series with summands expressed as values of a function in integers:

```
f(x)=x^3+2*x^2 + x - 1 ; g(x)=x^2-x+1
show(f.maxima_methods().divide(g))
```

Sage's output has the form  $[x + 3, 3x - 4]$ , where the first expression encodes the quotient and the second one the remainder. One may like to certify Sage's answer by adding the syntax `bool(g(x)*(x+3)+3*x-4==f(x))`, which returns True. Based now on (b) one has

$$\begin{aligned} \int \frac{f(x)}{g(x)} dx &= \int (x + 3) dx + \int \frac{3x - 4}{x^2 - x + 1} \\ &= \frac{x^2}{2} + 3x + \frac{3}{2} \ln(x^2 - x + 1) \\ &\quad - \frac{5}{\sqrt{3}} \arctan\left(\frac{2x - 1}{\sqrt{3}}\right) + C. \end{aligned}$$

Here, to compute the integral  $\ell = \int \frac{3x - 4}{x^2 - x + 1} dx$  you can apply the same method as 6.B.22 (since  $g(x)$  has only complex roots). In particular,

$$\begin{aligned} \ell &= \frac{3}{2} \int \frac{2x - 1}{x^2 - x + 1} dx - \frac{5}{2} \int \frac{dx}{(x - \frac{1}{2})^2 + (\frac{\sqrt{3}}{2})^2} \\ &= \frac{3}{2} \ln(x^2 - x + 1) - \frac{5}{\sqrt{3}} \arctan\left(\frac{2x - 1}{\sqrt{3}}\right) + c, \end{aligned}$$

for some constant  $c$ . This agrees with Sage's output for the command `integral((3*x-4)/(x^2-x+1),x)`.  $\square$

*Many integrals may initially appear complicated. In such cases, the power of the substitution method becomes evident. For instance, the substitution method often allows us to transform irrational functions into rational ones, which are easier to integrate.*

The tasks below focus on integrating irrational functions and serve as preparation for the material covered in the final section of this chapter, where more space is dedicated to this topic. But first, let us describe a few useful hints.

- Hint 1: This is about integrals of the form

$$\int f(x, \sqrt[p_1]{x}, \sqrt[p_2]{x}, \dots, \sqrt[p_k]{x}) dx$$

for certain numbers  $p_1, p_2, \dots, p_k \in \mathbb{N}$  and a rational function  $f$ . In this case we set  $\sqrt[n]{x} = t$ , i.e.,  $t^n = x$ , where  $n$  is the least common multiple of  $p_1, \dots, p_k$ . This substitution reduces the integrated function (integrand) to a rational function, which we can always integrate. When instead of the expressions  $\sqrt[p_j]{x}$  we have  $\sqrt[p_j]{ax + b}$  for all  $j = 1, \dots, k$ , where  $a, b \in \mathbb{R}$ , then set  $t^n = ax + b$ , where  $n$  occurs in the same vein as above.

- Hint 2: This is about integrals of the type

$$\int f\left(x, \sqrt[p_1]{\frac{ax+b}{cx+d}}, \sqrt[p_2]{\frac{ax+b}{cx+d}}, \dots, \sqrt[p_k]{\frac{ax+b}{cx+d}}\right) dx,$$

where the values  $a, b, c, d \in \mathbb{R}$  are such that  $ad - bc \neq 0$ . In this case we set  $t^n = \frac{ax+b}{cx+d}$ , where again  $n$  is the least

INTEGRAL CRITERION

**Theorem.** Let  $\sum_{n=1}^{\infty} f(n)$  be a series such that the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is positive and nonincreasing on the interval  $[1, \infty)$ . Then this series converges if and only if the integral

$$\int_1^{\infty} f(x) dx.$$

converges.

**PROOF.** If the integral is interpreted as the area of a region under the curve, the criterion is clear.



Indeed, notice that the given series diverges or converges if and only if the same is true for the same series without the first summand. Moreover, by the monotonicity of  $f(x)$ , there are the following estimates

$$\tilde{s}_k = \sum_{n=2}^k f(n) \leq \int_1^k f(x) dx \leq \sum_{n=1}^{\infty} f(n) = s_k,$$

because  $\tilde{s}_k$  is the lower sum of the Riemann integral while  $s_k$  is the upper sum. Thus, the integral converges if and only if the series does, as expected.  $\square$

**6.2.17. New acquisitions to the ZOO.** As a matter of fact,



primitive functions (Newton integrals) are very rarely described in terms of known elementary functions. Indeed, nearly all continuous functions lead to integrals which we cannot express in this way. Functions obtained by integration often appear in applications. Many of them have names and there are efficient methods how to approximate them numerically (we shall come to this point briefly in 6.2.23 below). Such functions either appear as the primitive functions or they are given by definite integrals dependent on further parameter(s). We present just a few (important) examples now.

In the methods of signal processing, a very important function is the so called *sinc function*:

$$(1) \quad \text{sinc}(x) = \frac{\sin(x)}{x}.$$

We shall meet the sinc function in our discussion of Fourier transform in 7.2.6. Check yourself that it is a smooth function with limit values

$$f(0) = 1, f''(0) = -\frac{1}{3}, \dots, f^{(2k)}(0) = (-1)^k \frac{1}{2k+1},$$

whilst all  $f^{(2k+1)}(0)$  vanish. This is easily seen from the Taylor expansion of the sine function (try to verify this by computing the derivatives directly - a good exercise on multiple use of L'Hospital even for very small orders). The even function sinc has its absolute maximum at the point  $x = 0$  and many local maxima and minima, with inflexions between them. It oscillates with a fast decreasing amplitude as  $x$  approaches infinity. The  $x$ -axis is the asymptot for both infinities.

common multiple of  $p_1, \dots, p_k$ . This converts the integrand to a rational function, as well.

**6.B.30.** Determine the integrals  $\int f(x) dx$ , where

(a)  $f(x) = \frac{1}{\sqrt{x^3 + \sqrt[5]{x^7}}}$ , with  $x > 0$ ;

(b)  $f(x) = \frac{x+1}{\sqrt[3]{3x+1}}$ , with  $x \neq -\frac{1}{3}$ ;

(c)  $f(x) = \frac{1}{x} \sqrt{\frac{x+1}{x-1}}$ , with  $x \in \mathbb{R} \setminus [-1, 1]$ .

**Solution.** (a) For all  $x > 0$  we have

$$f(x) = \frac{1}{\sqrt{x^3 + \sqrt[5]{x^7}}} = \frac{1}{\sqrt{x^2 \cdot x + \sqrt[5]{x^5 \cdot x^2}}} = \frac{1}{x(\sqrt{x + \sqrt[5]{x^2}})}.$$

The least common multiply of 2, 5 is 10. Hence, according to the first hint above to compute  $\int f(x) dx$  we set  $\sqrt[10]{x} = t$ , that is,  $t^{10} = x$  with  $10t^9 dt = dx$ . Then it is easy to see that  $\sqrt{x} = t^5$  and  $\sqrt[5]{x^2} = t^4$ , thus

$$\begin{aligned} \int f(x) dx &= \int \frac{dx}{x(\sqrt{x + \sqrt[5]{x^2}})} = \int \frac{10t^9}{t^{10}(t^5 + t^4)} dt \\ &= 10 \int \frac{dt}{t^6 + t^5}. \end{aligned}$$

Now we see that the function  $Q(t) = \frac{1}{t^6+t^5} = \frac{1}{t^5(1+t)}$  admits the following decomposition into simple fractions:

$$Q(t) = -\frac{1}{t+1} + \frac{1}{t} - \frac{1}{t^2} + \frac{1}{t^3} - \frac{1}{t^4} + \frac{1}{t^5}.$$

You can verify this expression by Sage and the command

```
var("t"); show((1/(t^6+t^5)).partial_fraction())
```

Thus

$$\begin{aligned} 10 \int Q(t) dt &= 10 \int \left( -\frac{1}{t+1} + \frac{1}{t} - \frac{1}{t^2} + \frac{1}{t^3} - \frac{1}{t^4} + \frac{1}{t^5} \right) dt \\ &= 10 \left( -\ln(t+1) + \ln(t) + \frac{1}{t} - \frac{1}{2t^2} + \frac{1}{3t^3} - \frac{1}{4t^4} \right) + C \\ &= \ln \frac{x}{(1 + \sqrt[10]{x})^{10}} + \frac{10}{\sqrt[10]{x}} - \frac{5}{\sqrt[5]{x}} + \frac{10}{3\sqrt[10]{x^3}} - \frac{5}{2\sqrt[5]{x^2}} + C \end{aligned}$$

for some constant  $C$ .

(b) Set  $t = \sqrt[3]{3x+1}$ , i.e.,  $t^3 = 3x+1$  and  $dx = t^2 dt$ . Then

$$\begin{aligned} \int \frac{x+1}{\sqrt[3]{3x+1}} dx &= \int \frac{\frac{t^3-1}{3} + 1}{t} t^2 dt = \int \frac{t^3-1+3}{3} t dt \\ &= \frac{1}{3} \int (t^4 + 2t) dt = \frac{1}{3} \left( \frac{t^5}{5} + t^2 \right) + C \\ &= \frac{1}{15} (3x+1)^{\frac{5}{3}} + \frac{1}{3} (3x+1)^{\frac{2}{3}} + C. \end{aligned}$$

Remember always to use Sage to verify your computations. For instance, one may find difficult to simplify the result of the back substitution  $t = \sqrt[3]{3x+1}$  in the expression  $\frac{1}{3} \left( \frac{t^5}{5} + t^2 \right)$ . There are many different ways to implement this, and here we present one based on the function subs.

GAMMA, SI, AND MORE

The sine integral function is defined by

(2) 
$$\text{Si}(x) = \int_0^x \text{sinc}(t) dt.$$

Other important functions are *Fresnel's sine and cosine integrals*

(3) 
$$\text{FresnelS}(x) = \int_0^x \sin\left(\frac{1}{2}\pi t^2\right) dt$$

(4) 
$$\text{FresnelC}(x) = \int_0^x \cos\left(\frac{1}{2}\pi t^2\right) dt.$$

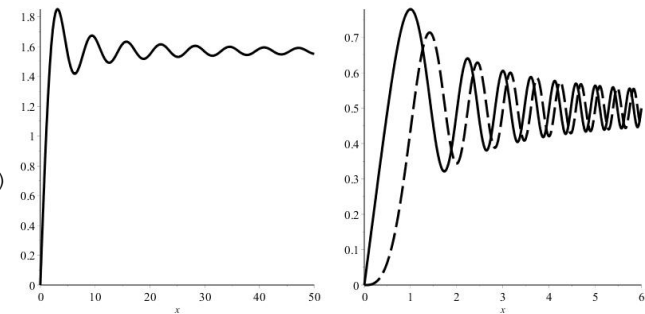
One of the most important mathematical functions ever is the *Gamma function*. It is defined for all positive real numbers  $z$  by

(5) 
$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt.$$

For all integers  $n \geq 0$ ,  $n! = \Gamma(n+1)$ .

The function  $\text{Si}(x)$  is shown in the left figure. Both Fresnel's functions are shown on the right.

The Gamma function is defined via an improper integral dependent on the parameter  $z$ . We shall provide some theory for such functions in the next part of this chapter. In particular, it can be proved that this function is analytic at all points  $0 < z \in \mathbb{R}$ .



For small  $z \in \mathbb{N}$ , we can evaluate:

$$\Gamma(1) = \int_0^\infty e^{-t} t^0 dt = [-e^{-t}]_0^\infty = 1$$

$$\Gamma(2) = \int_0^\infty e^{-t} t^1 dt = [-e^{-t} t]_0^\infty + \int_0^\infty e^{-t} dt = 1$$

$$\Gamma(3) = \int_0^\infty e^{-t} t^2 dt = 0 + 2 \int_0^\infty e^{-t} t dt = 0 + 2 = 2.$$

Integration by parts reveals immediately

$$\Gamma(z+1) = z\Gamma(z).$$

Hence for all positive integers  $n$  this function yields the value of the factorial:

$$\Gamma(n) = (n-1)!.$$

The following figure shows the behaviour of the functions  $f_1(x) = \ln(\Gamma(x+1))$ ,  $f_2(x) = \ln(\Gamma(x))$  together with the function  $x \ln x - x + 1$  (the dashed one, the dots are the actual values of  $\ln(n!)$ ).



```
var("t"); a(t)=(1/3)*((t^5/5) + t^2)
show(a.subs(t=(3*x+1)^(1/3)))
```

As for the given integral, just type

```
show(integral((x+1)/(3*x+1)^(1/3), x))
```

Observe that in our result there are still computations that can be done to arrive to a more solid expression, as  $\frac{1}{5}(3x+1)^{\frac{5}{3}}(x+2)$ . The same can be done in Sage, by adding in the last two show commands the function factor, that is, `show(a.subs(t=(3*x+1)^(1/3)).factor())` and `show(integral((x+1)/(3*x+1)^(1/3), x).factor())`, respectively. Check yourself Sage's outputs.

(c) In this case set  $t = \sqrt{\frac{x+1}{x-1}}$ , thus  $t^2 = \frac{x+1}{x-1}$  and  $x = \frac{t^2+1}{t^2-1}$ . Moreover, it is easy to see that  $dx = \frac{-4t}{(t^2-1)^2} dt$ . This substitution will convert the initial integrand to a rational functional, namely

$$\int \frac{1}{x} \sqrt{\frac{x+1}{x-1}} dx = \int \frac{t^2-1}{t^2+1} \frac{-4t^2}{(t^2-1)^2} dt = - \int \frac{4t^2}{(t^2+1)(t^2-1)} dt.$$

If  $Q(t) = -\frac{4t^2}{(t^2+1)(t^2-1)}$  then we see that

$$Q(t) = -2 \frac{2t^2}{(t^2+1)(t^2-1)} = -2 \left( \frac{(t^2+1) + (t^2-1)}{(t^2+1)(t^2-1)} \right) = -2 \left( \frac{1}{t^2+1} + \frac{1}{t^2-1} \right).$$

However,  $\frac{1}{t^2-1} = \frac{1}{2} \left( \frac{1}{t-1} - \frac{1}{t+1} \right)$ , and in total we get

$$Q(t) = -\frac{2}{t^2+1} - \frac{1}{t-1} + \frac{1}{t+1}.$$

One can recover this expression in Sage by the cell

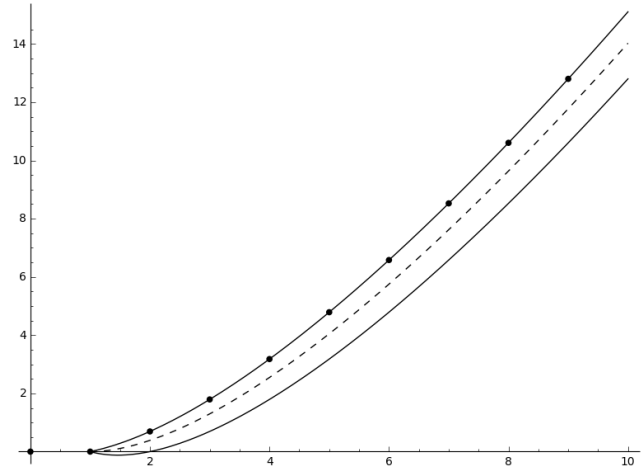
```
var("t")
(-(4*t^2)/((t^2-1)*(t^2+1))).partial_fraction()
```

Thus now we can rewrite our integral as

$$\begin{aligned} \int \frac{1}{x} \sqrt{\frac{x+1}{x-1}} dx &= \int \left( \frac{1}{t+1} - \frac{1}{t-1} - \frac{2}{t^2+1} \right) dt \\ &= \ln|t+1| - \ln|t-1| - 2 \arctan(t) + C \\ &= \ln \left| \frac{t+1}{t-1} \right| - 2 \arctan(t) + C \\ &= \ln \left| \frac{\sqrt{\frac{x+1}{x-1}} + 1}{\sqrt{\frac{x+1}{x-1}} - 1} \right| - 2 \arctan \left( \sqrt{\frac{x+1}{x-1}} \right) + C. \end{aligned}$$

For a Sage confirmation type the usual

```
show(integral((1/x)*sqrt((x+1)/(x-1)), x))
```



The relatively good approximation should not be a surprise, because we may clearly write  $\ln(n!) = \sum_{k=1}^n \ln k$  and the latter expression can be approximated by the integral  $\int_1^n \ln x = n \ln n - n + 1$ . Hence we came close to a very important and famous formula approximating  $n!$  by  $n^n e^{1-n}$ .

STIRLING'S FORMULA

The asymptotic estimate for the factorial function is

$$\ln n! = n \ln n - n + 1 + O(\ln n)$$

There is the famous *Stirling's approximation* formula making this estimate much more precise

$$(6) \quad \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq e n^{n+\frac{1}{2}} e^{-n}.$$

The lower approximation is so good, that the function  $\sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n}$  would completely overlap with the black one on the picture above.

We shall not provide the proof of the Stirling formula here. In order to understand the qualitative behavior of such functions (e.g. their differentiability etc.) we need to understand the limit processes much better. Before diving into this in the next section, we introduce several direct applications of the Riemann integral.

6.2.18. Riemann measurable sets. The definition of the



Riemann integral is motivated by the concept of the area of rectangles in the plane with coordinates  $x$  and  $y$ . The definite integral  $\int_a^b$  is designed to correspond to the area of the region bounded by the  $x$  axis, the values of the function  $y = f(x)$  and boundary lines  $x = a, x = b$ . Moreover, the area of the region above the  $x$  axis is given with a positive sign, while values under the axis lead to a negative sign.

From geometry, the length of an interval on the real line, and the area of a parallelogram determined by two vectors in the plane are basic concepts. This extends to the area of a parallelepiped in Euclidean vector space  $\mathbb{R}^n$ . The areas/volumes of other subsets are yet to be defined. For some simple objects like triangles, polygons and polyhedrons, their area is given naturally by the generally expected properties of area

□

Recall from 6.2.2 that for an arbitrary function  $f$  that is continuous and bounded on a bounded interval  $(a, b)$ , it holds the so called Newton-Leibniz formula, given by



$$\int_a^b f(x) dx = [F(x)]_a^b := \lim_{x \rightarrow b^-} F(x) - \lim_{x \rightarrow a^+} F(x). \quad (\star)$$

Here, as usual  $F'(x) = f(x)$  is the primitive function of  $f$ , with  $x \in (a, b)$ . Under the given conditions, the function  $F$  always exists and so do both the proper limits appearing in  $(\star)$ . Therefore, to compute the definite integral one just needs to find the antiderivative and determine the respective one-sided limits. We also recall that the computation of definite integrals via Sage is simple, and relies on the command `integral(f, x, a, b)`, or `f.integral(x, a, b)`, see below for examples.

**6.B.31.** Compute  $\int_{\frac{\pi}{6}}^{\frac{\pi}{3}} \tan^2(x) dx$  and  $\int_0^{\frac{\pi}{4}} \frac{x}{\cos^2(x)} dx$ . ○

**6.B.32.** Compute the integrals given below by hand, and next present a confirmation by Sage.

(a)  $\int_0^1 \frac{x}{\sqrt{1-x^2}} dx$ , (b)  $\int_1^2 \frac{1}{\sqrt{x^2-1}} dx$ . ○

**6.B.33.** Compute the integral  $\int_{\pi/4}^{\pi} \sin(3x) \cos(x) dx$  in Sage. Next provide a formal computation. ○

**6.B.34.** Prove the following inequalities:

(a)  $\frac{\sqrt{2}}{20} \leq \int_0^1 \frac{x^9}{\sqrt{1+x}} dx \leq \frac{1}{10}$ ;

(b)  $1 < \int_0^{\pi/2} \frac{\sin(x)}{x} dx < \frac{\pi}{2}$ . ○

**6.B.35.** Let  $f : [-a, a] \rightarrow \mathbb{R}$  be a continuous function,  $a \in \mathbb{R}$ .

(a) If  $f$  is even show that  $\int_{-a}^a f(x) dx = 2 \int_0^a f(x) dx$ .

(b) If  $f$  is odd show that  $\int_{-a}^a f(x) dx = 0$ . ○

**6.B.36.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function which is periodic with period  $T > 0$ .

(1) Show that the integral  $\int_a^{a+T} f(x) dx$  has the same value for every real number  $a$ , in particular,

$$\int_a^{a+T} f(x) dx = \int_0^T f(x) dx.$$

(2) More in general, show that

$$\int_a^{a+nT} f(x) dx = n \int_0^T f(x) dx,$$

(invariance with respect to Euclidean motions and additivity with respect to finite union of disjoint objects). We shall come to conceptual definitions in the next two chapters, mainly via generalizing the Riemann integral. Now we can start with several answers based on the univariate calculus.

We start with the question how to measure the “volume” of one-dimensional subsets.

We say that the subset  $A \subset \mathbb{R}$  is (Riemann) measurable, if the function  $\chi : \mathbb{R} \rightarrow \mathbb{R}$

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

is Riemann integrable. That is, the (improper) integral

$$m(A) = \int_{-\infty}^{\infty} \chi_A(x) dx$$

exists (the finiteness of its value doesn’t matter). The function  $\chi_A$  is called the characteristic function of the set  $A$ , the value  $m(A)$  is called the Riemann measure of the set  $A$ . Notice that for an interval  $A = [a, b]$  this yields the value

$$\int_{-\infty}^{\infty} \chi_A(x) dx = \int_a^b dx = b - a,$$

just as expected.

The elementary properties of the Riemann integral imply that this definition of “size” has reasonable properties. The measure of a union of finitely many measurable pairwise disjoint subsets is the sum of their measures. In particular, every finite set  $A$  has zero measure.

If instead we choose a countable union, this property is no longer true. For example, consider the set  $\mathbb{Q}$  of all rational numbers as the union of one-element subsets. While every set containing only finitely many points has a zero measure by our definition, the characteristic function  $\chi_{\mathbb{Q}}$  is not Riemann integrable over any finite interval  $[a, b]$ . This is an essential flaw of the Riemann approach and we shall comment on how to improve it in the end of this chapter.

Notice, the upper Riemann integral of the characteristic set  $\chi_A$  corresponds to the infimum of the sums of lengths of finitely many intervals, by which we can cover the given set  $A$ . The lower integral is the supremum of the sums of lengths of finitely many disjoint intervals that can be embedded into the set  $A$ . We shall proceed in the same way in higher dimensions when defining the Jordan measure. For now, just remark that the area of a plane figure bounded by a graph of a function in the way described above is consistent with expectations and we are going to deduce some straightforward consequences for special higher dimensional concepts.

**6.2.19. Length of a curve.** The Riemann integral can be effectively used to compute the length of a curve in multidimensional Euclidean vector space  $\mathbb{R}^n$ . For the sake of simplicity, we deal with a curve in  $\mathbb{R}^2$  with coordinates  $x, y$ . Suppose a parametric description of a curve  $F : \mathbb{R} \rightarrow \mathbb{R}^2$ ,



$$F(t) = [f(t), g(t)]$$

for any  $a \in \mathbb{R}$  and  $n \in \mathbb{Z}$ .

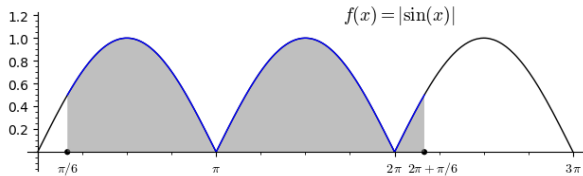
**6.B.37.** Use 6.B.36, to compute the following integrals:

- (a)  $\int_{\pi/6}^{\pi/6+2\pi} |\sin(x)| dx$ ;  
 (b)  $\int_{2\pi}^{4\pi} |\sin(x)| dx$ ;

**Solution.** (a) The function  $f(x) = |\sin(x)|$  is periodic with period  $T = \pi > 0$ . To confirm this claim in Sage use the `bool` command, as usual, i.e.,

```
f(x)=abs(sin(x)); bool(f(x+pi)==f(x))
```

The first integral corresponds to the grey area in the following figure:

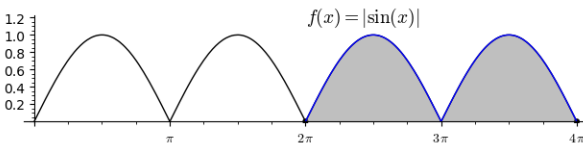


An application of the formula given in the second part of 6.B.36 gives

$$\begin{aligned} \int_{\pi/6}^{\pi/6+2\pi} |\sin(x)| dx &= \int_0^{2\pi} |\sin(x)| dx \\ &= 2 \int_0^{\pi} |\sin(x)| dx = 2 \int_0^{\pi} \sin(x) dx \\ &= 2[-\cos(x)]_0^{\pi} = 2 \cdot 2 = 4. \end{aligned}$$

To confirm this computation in Sage add in the previous cell the syntax `integral(f(x), x, pi/6, 2 * pi + pi/6)`.

(b) Since  $\int_0^{\pi} |\sin(x)| dx = 2$ , the second integral is also 4, see the figure given for this case:



Here is a formal computation:

$$\begin{aligned} \int_{2\pi}^{4\pi} |\sin(x)| dx &= \int_0^{4\pi} |\sin(x)| dx - \int_0^{2\pi} |\sin(x)| dx \\ &= 4 \int_0^{\pi} |\sin(x)| dx - 2 \int_0^{\pi} |\sin(x)| dx \\ &= (4 - 2) \int_0^{\pi} |\sin(x)| dx = 2 \cdot 2 = 4. \end{aligned}$$

Once more, to confirm the result via Sage just type `integral(abs(sin(x)), x, 2 * pi, 4 * pi)`.  $\square$

*According to the fundamental theorem of integral calculus, whenever  $f : A \subseteq \mathbb{R} \rightarrow \mathbb{R}$  is a continuous function and*

$\circ$  is given. Look at it as a trajectory of a movement. Assume that  $f(t)$  and  $g(t)$  have piece-wise continuous derivatives.

By differentiating the map  $F(t)$  we obtain vectors corresponding to the speed of the movement along this trajectory. Hence the total length of the curve (i.e. distance traveled over time between the values  $t = a, t = b$ ) is given by the integral over the interval  $[a, b]$ , with the integrated function  $h(t)$  being the length of the vectors  $F'(t)$ . Therefore the length  $s$  is given by the formula

$$s = \int_a^b h(t) dt = \int_a^b \sqrt{(f'(t))^2 + (g'(t))^2} dt.$$

The result can be seen intuitively as a corollary of Pythagoras' theorem: the linear increment  $\Delta s$  of the length of the curve corresponding to the increment  $\Delta t$  of variable  $t$  is given by the proportion in the orthogonal triangle and thus at the level of differentials

$$ds = \sqrt{(g'(t))^2 + (f'(t))^2} dt.$$

In the special case when the curve is the graph of a function  $y = f(x)$  between points  $a < b$ , we obtain

$$s = \int_a^b \sqrt{1 + (f'(x))^2} dx$$

and at the level of differentials,

$$ds = \sqrt{1 + (f'(x))^2} dx,$$

just as expected.

As an example, we calculate the circumference of the unit circle as twice the integral of the function  $y = \sqrt{1 - x^2}$  over  $[-1, 1]$ . We know that the result is  $2\pi$ , because  $\pi$  is defined in this way.

$$\begin{aligned} s &= 2 \int_{-1}^1 \sqrt{1 + (y')^2} dx = 2 \int_{-1}^1 \sqrt{1 + \frac{x^2}{1 - x^2}} dx \\ &= 2 \int_{-1}^1 \frac{1}{\sqrt{1 - x^2}} dx = 2[\arcsin x]_{-1}^1 = 2\pi. \end{aligned}$$

If we instead use

$$y = \sqrt{r^2 - x^2} = r\sqrt{1 - (x/r)^2}$$

and bounds  $[-r, r]$  in the previous calculation, by substituting  $x = rt$  we obtain the circumference of the circle with radius  $r$ :

$$\begin{aligned} s(r) &= 2 \int_{-r}^r \sqrt{1 + \frac{(x/r)^2}{1 - (x/r)^2}} dx = 2 \int_{-1}^1 \frac{r}{\sqrt{1 - t^2}} dt \\ &= 2r[\arcsin x]_{-1}^1 = 2\pi r. \end{aligned}$$

The result is of course well known from elementary geometry. Nevertheless, by using integral calculus, we derive the important fact that the length of a circle is linearly dependent on its diameter  $2r$ . The number  $\pi$  is exactly the ratio, appearing in this dependency.

$a \in A$ , the function

$$F(x) = \int_a^x f(t) dt, \quad x \in A,$$

satisfies  $F'(x) = f(x)$  for any  $x \in A$ . Therefore  $F$  is a primitive of  $f$  on  $A$ . This function plays a key role in integral calculus and in many cases can be used to simplify our computations. The next series of tasks are based on this scheme and will help us to master the fundamental theorem of calculus. Further similar exercises are presented in Section D.

**6.B.38.** (a) Find the derivative of the function  $F(x) = \int_0^x t^5 \ln(t+1) dt$ , with  $x \in (-1, 1)$ .

(b) If  $F(x) = \int_1^x (t^2 \sin(t) + 4 \cos(4t)) dt$ , with  $x > 0$ , show that  $\lim_{x \rightarrow \infty} g(x) = 1$ , where  $g(x) := \frac{F'(x)}{x^2+2}$ .  $\circ$

**6.B.39. Local extremes.** Find the local extremes of the function

$$F(x) = \int_0^x \frac{\sin(t)}{t} dt, \quad \text{with } x > 0.$$

**Solution.** For all  $x > 0$  we have  $F'(x) = \frac{\sin(x)}{x}$ . Thus, the critical points of  $F$ , i.e., the solutions of the equation  $F'(x) = 0$  have the form  $x = k\pi$  for  $k \in \mathbb{Z}_+$ , and they all provide local extremes of  $F$ . In particular, the second derivative of  $F$  has the form

$$F''(x) = \left( \frac{\sin(x)}{x} \right)' = \frac{x \cos(x) - \sin(x)}{x^2}, \quad x > 0,$$

and we see that  $F''(k\pi) = \frac{k\pi \cdot \cos(k\pi)}{(k\pi)^2} = \frac{(-1)^k}{k\pi}$ . Thus

- $F''(k\pi) < 0$  if  $k$  is odd  $\implies F$  attains a local maximum at  $(2m+1)\pi$ , with  $m \in \mathbb{N}$ .
- $F''(k\pi) > 0$  if  $k$  is even  $\implies F$  attains a local minimum at  $(2m)\pi$ , with  $m \in \mathbb{Z}_+$ .  $\square$

**6.B.40.** Find the third-order Taylor expansion of the function

$$f(x) = \int_0^x \frac{1}{1+t^2} dt, \quad x \in \mathbb{R},$$

around the point  $a = 0$ , both by hand and by Sage.

**Solution.** Recall by 6.A.8 that we want to determine the polynomial

$$T_0^3 f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f^{(3)}(0)}{6}x^3. \quad (b)$$

Obviously,  $f(0) = 0$  and

$$f'(x) = \frac{1}{1+x^2}, \implies f'(0) = 1,$$

$$f''(x) = \left( \frac{1}{1+x^2} \right)' = \frac{-2x}{(1+x^2)^2}, \implies f''(0) = 0,$$

$$f^{(3)}(x) = \left( \frac{-2x}{(1+x^2)^2} \right)' = \frac{2(3x^2-1)}{(1+x^2)^3}, \implies f^{(3)}(0) = -2.$$

**6.2.20. Areas and volumes.** The Riemann integral can be used to compute areas or volumes of shapes defined by a graph of a function. As an example, calculate the area of a circle with radius  $r$ . The quarter-circle bounded by the function  $\sqrt{r^2-x^2}$  for  $0 \leq x \leq r$  determines one quarter of the area. Use the substitution  $x = r \sin t$ ,  $dx = r \cos t dt$  (cf. the corollary for  $I_2$  in the paragraph 6.2.6) to obtain by symmetry

$$\begin{aligned} a(r) &= 4 \int_0^r \sqrt{r^2-x^2} dx = 4r^2 \int_0^{\pi/2} \cos^2 t dt \\ &= 4r^2 \int_0^{\pi/2} \sin^2 t dt = \frac{1}{2} 4r^2 \int_0^{\pi/2} \cos^2 t + \sin^2 t dt \\ &= 2r^2 \int_0^{\pi/2} dt = \pi r^2. \end{aligned}$$

It is worth noticing that this well known formula is derived from the principles of integral calculus. The area of a circle is not only proportional to the square of the radius, but this proportion is again given by the constant  $\pi$ .

Notice the ratio of the area to the perimeter of a circle.

$$\frac{\pi r^2}{2\pi r} = \frac{1}{2}r.$$

The square with the same area has the side of length  $\sqrt{\pi r}$  and therefore its perimeter is  $4\sqrt{\pi r}$ . Hence the perimeter of a square with the area of the unit circle is  $4\sqrt{\pi}$ , compared to the perimeter  $2\pi$  of the unit circle, which is about 0.8 less. It can be shown that in fact the circle is the shape with the smallest perimeter among all with the same area. We derive such results in comments about the calculus of variations in chapter 9.

Another analogy of this approach is the computation of the volume or the surface area of a *solid of revolution*. Such a set in  $\mathbb{R}^3$  is defined by plotting the graph of a function  $y = f(x)$  (for  $x$  in an interval  $[a, b]$ ) in the plane  $xy$  and rotating this plane around the  $x$  axis. This is exactly what happens when producing pottery on a jigger – the hands shape the clay in the form of  $y = f(x)$ .

add appropriate picture here!

When computing the area of the surface, an increment  $\Delta x$  causes the area to increase by the multiple of the length  $\Delta s$  of the curve given by the graph of the function  $y = f(x)$  and the size of the circle with radius  $f(x)$ . Hence the surface area  $A(f)$  is computed by the formula

$$A(f) = 2\pi \int_a^b f(x) ds = 2\pi \int_a^b f(x) \sqrt{1+(f'(x))^2} dx,$$

where the differential  $ds$  is given by the increment on the length of curve  $y = f(x)$ , see above. If instead we determine the solid of revolution by its bound parametrized in the  $xy$  plain by a pair of functions  $[x(t), y(t)]$ , then the corresponding differential of the length  $s$  has the form  $ds = \sqrt{(x'(t))^2 + (y'(t))^2} dt$ . Thus we obtain

$$A = 2\pi \int_a^b y(t) \sqrt{(y'(t))^2 + (x'(t))^2} dt.$$

Thus, according to (b) we get  $T_0^3 f(x) = x - \frac{x^3}{3}, x \in \mathbb{R}$ .

In Sage, as in 6.A.9 one can use the command `taylor`, though now the variable  $t$  should be included as a symbolic variable and assume that  $x > 0$ :

```
var("x, t"); assume(x>0)
f(x)=integral(1/(1+t^2), t, 0, x)
T(x)=taylor(f(x), x, 0, 3); show(T(x))
```

Sage's output has the form  $-\frac{1}{3}x^3 + x$ . □

**6.B.41.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a differentiable function on  $\mathbb{R}$ , having continuous second derivative everywhere on  $\mathbb{R}$  and a local extremum at  $x_0 = 1$ . If for some constants  $\alpha, \beta \in \mathbb{R}$  with  $\alpha \neq 1$  we have the relation

$$\int_0^1 (x f''(x) + \alpha f'(x)) dx = \beta \quad (*)$$

and the graph  $C_f$  of  $f$  passes through the point  $P = [0, \beta] \in \mathbb{R}^2$ , compute  $f(1)$  in terms of  $\alpha$  and  $\beta$ .

**Solution.** By assumption,  $f$  is differentiable everywhere on  $\mathbb{R}$  and hence also at  $x_0 = 1$ , and moreover attains a local extremum at  $x_0$ , so we have  $f'(x_0) = f'(1) = 0$ . We also have  $f(0) = \beta$ , since  $P$  belongs to the graph of  $f$ . Thus, if  $A$  is the left-hand-side of  $(*)$ , by applying integration by parts we get

$$\begin{aligned} A &= \int_0^1 x f''(x) dx + \alpha \int_0^1 f'(x) dx = [x f'(x)]_0^1 \\ &\quad - \int_0^1 x' f'(x) + a [f(x)]_0^1 = f'(1) - [f(x)]_0^1 + a [f(x)]_0^1 \\ &= 0 - (f(1) - f(0)) + a(f(1) - f(0)) \\ &= (\alpha - 1)(f(1) - f(0)) = (\alpha - 1)(f(1) - \beta). \end{aligned}$$

Thus we should have  $(\alpha - 1)(f(1) - \beta) = \beta$  from where we get  $f(1) = \frac{\beta}{\alpha - 1} + \beta = \frac{\alpha\beta}{\alpha - 1}$ . □

**6.B.42.** Consider the function

$$F(x) = x \int_1^x f(t) dt + (1 - x) \int_0^x f(t) dt$$

with  $x \in [0, 1]$ , where  $f : [0, 1] \rightarrow \mathbb{R}$  is continuous function on  $\mathbb{R}$  with  $f(0)f(1) \neq 0$ . Show that there exists  $\xi \in (0, 1)$  with

$$f(\xi) = \int_0^1 f(t) dt. \quad (*)$$

**Solution.** The function  $F$  is differentiable on  $[0, 1]$  and it easy to see that it satisfies  $F(0) = 0 = F(1)$ . Therefore, by Rolle's theorem there exists  $\xi \in (0, 1)$  with  $F'(\xi) = 0$ . We will show that  $\xi$  satisfies  $(*)$ . In particular, a combination of the product rule with the fundamental theorem of calculus

When computing the volume of the same solid, then the increase of  $\Delta x$  causes the volume increase by a multiple of this increment and the area of the circle with radius  $f(x)$ . Hence it is given by the formula

$$V(f) = \pi \int_a^b (f(x))^2 dx.$$

As an example of using the formulas for surface and volume, we derive the well known formulas for the surface of the sphere and volume of the ball with diameter  $r$ .

$$\begin{aligned} A_r &= 2\pi \int_{-r}^r r \sqrt{1 - (x/r)^2} \frac{1}{\sqrt{1 - (x/r)^2}} dt \\ &= 2\pi r \int_{-r}^r dt = 4\pi r^2 \\ V_r &= \pi \int_{-r}^r (r^2 - x^2) dx \\ &= \pi \left[ r^2 x - \frac{1}{3} x^3 \right]_{-r}^r = \frac{4}{3} \pi r^3. \end{aligned}$$

Similarly to the circle, the ball is also the object which has the smallest surface among all with the given volume. That is the reason why small soap bubbles almost always assume this shape.

**6.2.21. Differential equations.** Theorem 6.2.9 can be understood in the following way. Given a continuous function  $f(x)$  on a bounded interval  $[a, b]$ , the set of all functions  $y$  of one variable  $x$  satisfying the equality



$$y' = f(x)$$

is given by the formula

$$y(x) = \int_a^x f(t) dt + C$$

with the constant  $C = y(a)$ . This is the simplest instance of *differential equations*. More generally, ordinary differential equations of first order are given as

$$y' = f(x, y),$$

where  $f(x, y)$  depends on two unknown variables  $x$  and  $y$ . A solution to this equation is a function  $y = y(x)$ , such that the equality is true upon substitution. Similarly, dependence on higher derivatives of  $y$  may be included.

We return to this concept in Chapter 8, see 8.3.2. For the present, we discuss one very special type of equations with *separated variables*

$$y' = f(x)g(y)$$

and add a few observations concerning analytic solutions. Rewrite the equation in terms of the differentials, cf. 6.1.11,

$$\frac{1}{g(y)} dy = f(x) dx.$$

Find the primitive functions on both sides to determine the unknown function  $y = y(x)$  implicitly.

(see 6.2.9), gives

$$\begin{aligned} F'(x) &= \left( x \int_1^x f(t) dt \right)' + \left( (1-x) \int_0^x \right)' \\ &= \int_1^x f(t) dt + x f(x) - \int_0^x f(t) dt + (1-x)f(x) \\ &= - \left( \int_0^x f(t) dt + \int_x^1 f(t) dt \right) + f(x) \\ &= - \int_0^1 f(t) dt + f(x). \end{aligned}$$

This expression combined with the equation  $F'(\xi) = 0$  yields the result.  $\square$

**6.B.43.** Let  $f : [0, 1] \rightarrow \mathbb{R}$  be a continuous function. Show that

$$\int_0^\pi x f(\sin(x)) dx = \frac{\pi}{2} \int_0^\pi f(\sin(x)) dx.$$

**Solution.** Set  $A = \int_0^\pi x f(\sin(x)) dx$ . The trick is the substitution  $u = \pi - x$  with  $du = -dx$ . For  $x = 0$  we have  $u = \pi$  and for  $x = \pi$  we have  $u = 0$ . Recall also that

$$\begin{aligned} \sin(\pi - u) &= \sin(\pi) \cos(u) - \sin(u) \cos(\pi) \\ &= 0 - (-1) \sin(u) = \sin(u), \end{aligned}$$

for all  $u \in \mathbb{R}$ . Thus

$$\begin{aligned} A &= - \int_\pi^0 (\pi - u) f(\sin(\pi - u)) du \\ &= \int_0^\pi (\pi - u) f(\sin(u)) du \\ &= \pi \int_0^\pi f(\sin(u)) du - \int_0^\pi u f(\sin(u)) du, \end{aligned}$$

and the result follows easily.  $\square$

*Along the proof of the fundamental theorem of integral calculus we used the notion of “uniformly continuous functions”, see 6.2.11. Uniform continuity is a stronger condition than continuity, since the real number  $\delta$  in the relative definition depends only on  $\varepsilon$ . Thus, every uniformly continuous function defined on a subset  $A \subseteq \mathbb{R}$  is also continuous, but the converse is not true (think for example of the hyperbola  $1/x$  on  $(0, 1)$ ). Let us present an alternative way to think about uniformly continuous functions and discuss some examples.*



**6.B.44.** Show that a function  $f : A \subseteq \mathbb{R} \rightarrow \mathbb{R}$  is uniformly continuous if and only if for any two sequences  $(x_n), (y_n)$  in  $A$  with  $x_n - y_n \rightarrow 0$ , we also have  $f(x_n) - f(y_n) \rightarrow 0$ , as  $n \rightarrow \infty$ .

**Solution.** Let us prove the one direction and leave the converse as an exercise. So, let  $f : A \subseteq \mathbb{R} \rightarrow \mathbb{R}$  be a uniformly continuous function. Then, for every  $\varepsilon > 0$  there exists  $\delta > 0$  such that for all  $x, y \in A$  satisfying  $|x - y| < \delta$  the inequality  $|f(x) - f(y)| < \varepsilon$  holds. Let  $(x_n), (y_n)$  be sequences on  $A$

Indeed, if  $G(y)$  and  $F(x)$  are the primitive functions with  $G'(y) = \frac{1}{g(y)}$  and  $F'(x) = f(x)$ , and  $y(x)$  satisfies  $G(y(x)) = F(x)$ , then differentiating both sides with respect to  $x$  yields

$$0 = G'(y(x))y'(x) - F'(x) = \frac{y'(x)}{g(y(x))} - f(x)$$

as expected. Of course, it is necessary to be careful with the values  $y$  for which  $g(y) = 0$ , which need to be discussed separately.

For example, the equation  $y' = y$  leads to the implicit definition

$$\ln |y| = x + C,$$

which for positive  $y$  provides  $y(x) = D e^x$  with positive constant  $D$ , the constant solution  $y = 0$  corresponds to  $D = 0$ . Negative values of  $y$  correspond to negative constants  $D$  in the same expression.

If  $y(0) = 1$ , we recover the exponential  $y(x) = e^x$ .

**6.2.22. Analytic solutions.** As we know, the power series are differentiated and integrated term by term, thus the solution  $y(x)$  to the equation  $y' = f(x)$  with a known analytic function  $f(x) = \sum_{n=0}^\infty a_n x^n$  must be



$$y(x) = \sum_{n=0}^\infty \frac{1}{n+1} a_n x^{n+1} + y(0),$$

where  $y(0)$  is the free integration constant. The solution is defined on the convergence domain of the power series (which has to be same as that of  $f$  by the lim sup formula). Of course we might use series centered in other points  $x_0$  if prescribing the initial value  $y(x_0)$ . (We shall prove much later, that actually there is always the unique solution with the given initial prescribed value  $y(0)$  in Chapter 8.)

The latter equation  $y' = y$  had the analytic solution  $e^x$ , too. Let us consider the general case of this type, i.e. equations of the form

$$(1) \quad y' = f(y)$$

with an analytic right-hand side  $f(y)$ . Given the initial condition  $y(x_0) = y_0$ , straightforward differentiation with the help of the chain rule and the equation (1) shows

$$\begin{aligned} y'(x_0) &= f(y_0) \\ y''(x_0) &= f'(y)y'|_{x=x_0} = f'(y)f(y)|_{x=x_0} = f'(y_0)f(y_0) \\ y'''(x_0) &= (f''(y)y'f(y) + f'(y)f'(y)y')|_{x=x_0} \\ &= f''(y_0)(f(y_0))^2 + (f'(y_0))^2 f(y_0) \\ &\vdots \end{aligned}$$

Two crucial observations are due here. First, giving the initial condition  $y(x_0) = y_0$ , all derivatives  $y^{(k)}(x_0)$  are given at this point by the equation. Thus, if an analytic solution exists, we know it explicitly. So we have to focus on the convergence of the known formal expression of the series  $y(x) = \sum_{n=0}^\infty \frac{1}{n!} y^{(n)}(x_0)(x - x_0)^n$  and we arrive at the theorem below.

with  $x_n - y_n \rightarrow 0$ , and let  $\varepsilon > 0$  be given. By the definition of a convergent sequence this means that there exists some natural  $n_0$  such that  $|x_n - y_n| < \varepsilon$  for all  $n \geq n_0$ . But then we also get  $|f(x_n) - f(y_n)| < \varepsilon$  for all  $n \geq n_0$ , which is equivalent to say that  $\lim_{n \rightarrow \infty} (f(x_n) - f(y_n)) = 0$ .  $\square$

**6.B.45.** Let  $f : A \rightarrow \mathbb{R}$  be a continuous function defined on subset  $A$  of  $\mathbb{R}$ . Provide examples verifying that:

- (a) If  $A$  is not closed, then  $f$  may not be uniformly continuous on  $A$ ;
- (b) If  $A$  is not bounded, then  $f$  may not be uniformly continuous on  $A$ .  $\circ$

**6.B.46.** Based on 6.B.44, show that the hyperbola  $f(x) = 1/x$  with  $x \in (0, 1)$  is not uniformly continuous.  $\circ$

**6.B.47.** Show that the sine function  $\sin(x)$  is uniformly continuous on  $\mathbb{R}$ .

**Solution.** First we will use the Lagrange’s mean value theorem (see 5.3.9) to show that  $|\sin(x) - \sin(y)| \leq |x - y|$ , for any  $x, y \in \mathbb{R}$ . If  $x = y$  we have nothing to prove, so assume that  $x \neq y$ . Without loss of generality we may also assume that  $x < y$ . The sine function  $\sin(x)$  satisfies the conditions of the Lagrange’s mean value theorem on the interval  $[x, y]$ . Thus, there exists some  $\xi \in (x, y)$  with

$$\frac{\sin(x) - \sin(y)}{x - y} = \sin'(\xi) = \cos(\xi).$$

Since  $|\cos(\xi)| \leq 1$  we get  $\left| \frac{\sin(x) - \sin(y)}{x - y} \right| \leq 1$ , which implies that  $|\sin(x) - \sin(y)| \leq |x - y|$ , for any  $x, y \in \mathbb{R}$ .

Let us now prove that  $\sin(x)$  is uniformly continuous on  $\mathbb{R}$ . Given some  $\varepsilon > 0$  take  $\delta = \varepsilon$ . If  $x, y \in \mathbb{R}$  satisfy  $|x - y| < \delta$ , by the previous inequality we will have  $|\sin(x) - \sin(y)| \leq |x - y| < \delta = \varepsilon$ , which proves the claim.  $\square$

**6.B.48.** Show that the cos function  $\cos(x)$  is uniformly continuous on  $\mathbb{R}$ .  $\circ$

**6.B.49.** Let  $f : A \rightarrow \mathbb{R}$  be a continuous function defined on a subset  $A \subseteq \mathbb{R}$ .

- (a) Show with an example that if  $(x_n)$  is a Cauchy sequence on  $A$ , then the sequence  $(f(x_n))$  may fail to be Cauchy.
- (b) If  $f$  is in addition uniformly continuous on  $A$ , show that the sequence  $(f(x_n))$  is Cauchy for any Cauchy sequence  $(x_n)$  on  $A$ .  $\circ$

**6.B.50.** Which of the following functions is uniformly continuous on the given domain?

$f(x) = x^2$  on  $A = [0, 1]$ ,  $g(x) = \tan(x)$  on  $B = [0, \pi/2)$ ,  $h(x) = x^2$  on  $\mathbb{R}$ , and  $k(x) = x^3$  on  $\mathbb{R}$ .  $\circ$



In many cases we should compute integrals over an infinite interval, or integrals of functions containing a discontinuity, see 6.2.14. In such cases one speaks for “improper integrals”, which are in general definite integrals that cover an unbounded area. For convenience we summarize the following rules:

In its proof, the second observation will be most helpful: the expressions for the derivatives  $y^{(n)}$  are universal polynomials  $P_n$

$$y^{(n)}(x) = P_n(f(x), f'(x), \dots, f^{(n-1)}(x))$$

in the derivatives of the function  $f$ , all with non-negative coefficients and independent of the particular equation.<sup>5</sup>

CAUCHY-KOVALEVSKAYA THEOREM IN DIMENSION ONE

**Theorem.** Assume  $f(y)$  is a real analytic function convergent on the interval  $(x_0 - a, x_0 + a) \subset \mathbb{R}$  and consider the differential equation (1) with the condition  $y(x_0) = y_0$ . Then the formal power series  $y(x) = \sum_{n=0}^{\infty} \frac{1}{n!} y^{(n)}(x_0)(x - x_0)^n$  converges on a neighborhood of  $x_0$  and provides the solution to (1) satisfying the initial condition.

**PROOF.** The second observation above suggests how to prove the convergence of the “candidate series” similarly as we proved the convergence of power series in general, i.e. by finding another converging series whose partial sums will bound our’s from above. This was the original Cauchy’s approach to this theorem and we talk about the *method of majorants*.

Without loss of generality we shall fix  $x_0 = 0$  and  $y(0) = 0$  (we may always use shifted quantities  $z = y - y_0$  and  $t = x - x_0$  to transform the general case). Assume we can find another analytic function  $g(x) = \sum_{n=0}^{\infty} \frac{1}{n!} b_n x^n$  with all  $b_n = g^{(n)} \geq 0$ , i.e.  $g$  has got all derivatives non-negative at the origin, such that  $g^{(n)}(0) \geq |f^{(n)}(0)|$  for all  $n$ .

Now, replace  $f$  in the equation (1) by  $g$  and write formal power series  $z(x) = \sum_{n=0}^{\infty} \frac{z^{(n)}}{n!} x^n$  for the potential solution of this equation as above. In particular, we deduce (recall the universal polynomials  $P_n$  have got non-negative coefficients)

$$\begin{aligned} z^{(n)}(0) &= P_n(g(z(0)), \dots, g^{(n-1)}(z(0))) \\ &\geq P_n(|f(y(0))|, \dots, |f^{(n-1)}(y(0))|) \geq |y^{(n)}(0)| \end{aligned}$$

and, consequently, convergence of  $z(x)$  will imply absolute convergence of  $y(x)$ , i.e. the claim of the Theorem. We try to find a majorant in the form of a geometric series.

Let us pick  $r > 0$ , smaller than the radius of convergence of  $f$ . Then obviously, there is a constant  $C > 0$  such that the derivatives  $a_n = f^{(n)}(0)$  satisfy  $|\frac{1}{n!} a_n r^n| \leq C$  for all  $n$ , i.e.  $|a_n| \leq C \frac{n!}{r^n}$  (the series would certainly not converge otherwise). We may recognize the derivatives of a geometric series and write

$$(2) \quad g(z) = C \sum_{n=0}^{\infty} \frac{z^n}{r^n} = C \frac{r}{r - z}$$

with derivatives  $g^{(n)}(0) = C \frac{n!}{r^n}$ .

Finally, we have to prove that the solution of the equation  $z' = g(z)$  is analytic. We can easily integrate this equation

<sup>5</sup>Although we shall not need the explicit formulae for these polynomials, they are well known under the name *Faà di Bruno’s formula*. In principle, they are direct generalization of the Leibniz rule to higher order derivatives.

(1)  $\int_a^{+\infty} f(x) dx = \lim_{c \rightarrow +\infty} \int_a^c f(x) dx$ ,  
 where  $f$  is assumed to be continuous on  $[a, +\infty)$ .

(2)  $\int_{-\infty}^b f(x) dx = \lim_{c \rightarrow -\infty} \int_c^b f(x) dx$ ,  
 where  $f$  is assumed to be continuous on  $(-\infty, b]$ .

(3)  $\int_{-\infty}^{+\infty} f(x) dx = \int_{-\infty}^0 f(x) dx + \int_0^{+\infty} f(x) dx$ ,  
 where  $f$  is assumed to be continuous on  $\mathbb{R}$ .

(4)  $\int_a^b f(x) dx = \lim_{t \rightarrow b^-} \int_a^t f(x) dx$ ,  
 where  $f$  is assumed to be continuous on  $[a, b)$ .

(5)  $\int_a^b f(x) dx = \lim_{t \rightarrow a^+} \int_t^b f(x) dx$ ,  
 where  $f$  is assumed to be continuous on  $(a, b]$ .

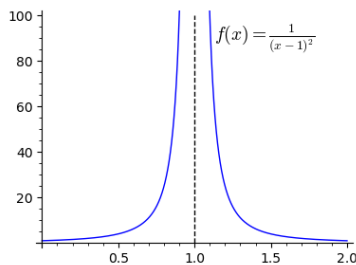
(6)  $\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx$ ,  
 where  $f$  is assumed to be continuous on  $[a, c) \cup (c, b]$ .

**6.B.51.** Show that the integral  $L = \int_0^2 \frac{1}{(x-1)^2} dx$  is divergent, i.e.,  $L = +\infty$ .

**Solution.** Observe that  $x_0 = 1$  is an improper point of the function  $f(x) = 1/(x-1)^2$  and

$$\lim_{x \rightarrow 1^+} f(x) = \lim_{x \rightarrow 1^-} f(x) = \lim_{x \rightarrow 1} f(x) = +\infty.$$

Hence the line  $x = 1$  is a vertical asymptote of  $f$ . However the integrand  $f$  is continuous at all other points, see also its graph below.



Already from the graph we understand that the statement should be true. To prove it we will extend the method presented in 6.2.14 for computing the integral  $I$  appearing there, in particular this improper integral corresponds to case (6) of those listed above. Hence we have

$$\begin{aligned} L &= \int_0^2 f(x) dx = \int_0^1 f(x) dx + \int_1^2 f(x) dx \\ &= \lim_{\delta \rightarrow 1^-} \int_0^\delta f(x) dx + \lim_{\varepsilon \rightarrow 1^+} \int_\varepsilon^2 f(x) dx. \end{aligned}$$

with separated variables directly. Written with the help of differentials,  $(r-z) dz = Cr dx$ . Thus, the implicit equation reads  $\frac{1}{2}(r-z)^2 = -Cr x + D$ , where the constant  $D$  is determined by  $z(0) = 0$ . Consequently  $D = \frac{1}{2}r^2$  and a simple computation reveals the solution of the implicit equation

$$z(x) = r \left( 1 \pm \sqrt{1 - \frac{2Cx}{r}} \right).$$

The option with the minus sign satisfies our initial condition. This clearly is an analytic function,  $g$  provides the requested majorant, and the proof is finished.  $\square$

**6.2.23. Numerical approximation of integration.** Just



as in paragraph 6.1.12, we use the Taylor expansion to propose simple approximations of integration. We deal with an integral  $I = \int_a^b f(x) dx$  of an analytic function  $f(x)$  and a uniform partition of the interval  $[a, b]$  using points  $a = x_0, x_1, \dots, x_n = b$  with distances  $x_i - x_{i-1} = h > 0$ . Denote the points in the middle of the intervals in the partitions by  $x_{i+1/2}$  and the values of the function at the points of the partition by  $f(x_i) = f_i$ .

Compute the contribution of one segment of the partition to the integral by the Taylor expansion (knowing that the power series might be integrated term by term). Integrate symmetrically around the middle values so that the derivatives of odd orders cancel each other out while integrating:

$$\begin{aligned} \int_{-h/2}^{h/2} f(x_{i+1/2} + t) dt &= \int_{-h/2}^{h/2} \left( \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x_{i+1/2}) t^n \right) dt \\ &= \sum_{k=0}^{\infty} \left( \int_{-h/2}^{h/2} \frac{1}{k!} f^{(k)}(x_{i+1/2}) t^k dt \right) \\ &= \sum_{k=0}^{\infty} \frac{h^{2k+1}}{2^{2k}(2k+1)!} f^{(2k)}(x_{i+1/2}). \end{aligned}$$

A simple numerical approximation of integration on one segment of the partition is the *trapezoidal rule*. This uses the area of a trapezoid given by the points  $[x_i, 0], [x_i, f_i], [x_{i+1}, 0], [x_{i+1}, f_{i+1}]$  for approximation. This area is

$$P_i = \frac{1}{2}(f_i + f_{i+1})h.$$

In total, the integral  $I$  is approximated by

$$I_{\text{trap}} = \sum_{i=0}^{n-1} P_i = \frac{h}{2}(f_0 + 2f_1 + \dots + 2f_{n-1} + f_n).$$

Compare  $I_{\text{trap}}$  to the exact value of  $I$  computed by contributions over individually segments of the partition. Express the values  $f_i$  by the middle values  $f_{i+1/2}$  and the derivatives  $f_{i+1/2}^{(k)}$  in the following way:

$$\begin{aligned} f_{i+1/2 \pm 1/2} &= f_{i+1/2} \pm \frac{h}{2} f'_{i+1/2} + \frac{h^2}{2!2^2} f''_{i+1/2} \\ &\quad \pm \frac{h^3}{3!2^3} f^{(3)}_{i+1/2} + \dots \end{aligned}$$



Now, obviously  $F(x) = -1/(x - 1)$  satisfies  $F'(x) = f(x)$  and hence is a primitive of  $f$ . Applying the fundamental theorem of integral calculus we obtain

$$\int_0^\delta f(x) dx = F(\delta) - F(0) = -\frac{\delta}{\delta - 1},$$

$$\int_\varepsilon^2 f(x) dx = F(2) - F(1 + \varepsilon) = -\frac{\varepsilon - 2}{\varepsilon - 1},$$

and hence

$$L = \lim_{\delta \rightarrow 1^-} \frac{\delta}{1 - \delta} + \lim_{\varepsilon \rightarrow 1^+} \frac{\varepsilon - 2}{1 - \varepsilon} = (+\infty) + (+\infty) = +\infty.$$

These computations can be done quickly in Sage, e.g., by the block

```
F(x)=-1/(x-1); var("delta, eps")
show((F(delta)-F(0)).factor())
show(lim((F(delta)-F(0)), delta=1, dir="-"))
show((F(2)-F(eps)).factor())
show(lim((F(2)-F(eps)), eps=1, dir="+"))
```

Notice the command `integral(1/(x - 1)2, x, 0, 2)` gives an error which confirms itself the divergence of  $L$ , check yourself the very end of Sage's output for this command.  $\square$

**6.B.52.** Show that  $\int_1^{+\infty} \frac{\arctan(x)}{x\sqrt{x}} dx$  is a finite real number.

○

In 6.2.17 one meets several new acquisitions to the ZOO. One of them is the Gamma function, defined for all positive real numbers  $z$  by

$$\Gamma(z) = \int_0^\infty e^{-x} x^{z-1} dx.$$

The Gamma function is well defined and analytic for  $z > 0$  (and more generally for complex numbers  $z$  with positive real part). Moreover,  $\Gamma$  converges for all  $z > 0$  and the next task shows that the Gamma function can be seen as an extension of the factorial function to real positive numbers.<sup>4</sup>

**6.B.53. An interpretation of the factorial.** (a) Prove the recurrence formula  $\Gamma(n + 1) = n\Gamma(n)$ ,  $n \in \mathbb{Z}_+$ . (b) Using (a) show that  $\Gamma(n + 1) = n!$ , for all naturals  $n$ . (c) For some  $a > 0$  prove that

$$\int_0^\infty e^{-ax} x^{n-1} dx = \frac{1}{a^n} \Gamma(n).$$

(d) Using the “Gaussian integral”  $\int_0^\infty e^{-u^2} du = \frac{\sqrt{\pi}}{2}$ , show that  $\Gamma(1/2) = \sqrt{\pi}$ .

<sup>4</sup>The gamma function was introduced by the Leonhard Euler (1707-1783) at 1729, as a natural extension of the factorial operation  $n!$  (in the sense that  $n$  was replaced by a real or a complex number).

Thus, the contribution  $P_i$  to the approximation is

$$P_i = \frac{1}{2}(f_i + f_{i+1})h = h \left( f_{i+1/2} + \frac{h^2}{2!2^2} f''(i+1/2) \right) + O(h^5).$$

Estimate the error  $\Delta_i = I - I_{\text{trap}}$  over one segment of the partition:

$$\Delta_i = h \left( f_{i+1/2} + \frac{h^2}{24} f''_{i+1/2} - f_{i+1/2} - \frac{h^2}{8} f''_{i+1/2} + O(h^4) \right)$$

$$= -\frac{h^3}{12} f''_{i+1/2} + O(h^5).$$

The total error is thus estimated as

$$|I - I_{\text{trap}}| = \frac{1}{12} n h^3 |f''| + n O(h^5)$$

$$= \frac{1}{12} (b - a) h^2 |f''| + O(h^4)$$

where  $|f''|$  represents an upper estimate for  $|f''(x)|$  of  $f$  over the interval of integration.

If the linear approximation of the function over the individual segments does not suffice, we can try approximations by quadratic polynomials. To do so, three values are always needed, so work with segments of the partition in pairs. Suppose  $n = 2m$  and consider  $x_i$  with odd indices. We choose

$$f_{i+1} = f(x_i + h) = f_i + \alpha_i h + \beta_i h^2$$

$$f_{i-1} = f(x_i - h) = f_i - \alpha_i h + \beta_i h^2$$

which implies

$$\beta_i = \frac{1}{2h^2} (f_{i+1} + f_{i-1} - 2f_i).$$

The approximation of the integral over two segments of the partition between  $x_{i-1}$  and  $x_{i+1}$  is now estimated by the expression (notice we integrate the quadratic polynomial with the requested values  $f_{i-1}$ ,  $f_i$ ,  $f_{i+1}$  in the points  $x_{i-1}$ ,  $x_i$ ,  $x_{i+1}$ , respectively. It is not necessary to know the constant  $\alpha_i$ )

$$P_i = \int_{-h}^h f_i + \alpha_i t + \beta_i t^2 dt = 2h f_i + \frac{2}{3} \beta_i h^3$$

$$= 2h f_i + \frac{2h}{6} (f_{i+1} + f_{i-1} - 2f_i)$$

$$= \frac{h}{3} (f_{i+1} + f_{i-1} + 4f_i).$$

This procedure is called *Simpson's rule*<sup>6</sup>. The entire integral is now approximated by

$$I_{\text{Simp}} = \frac{1}{3} h (f_0 + 4 \sum_{m=0}^{n-1} f_{2m+1} + 2 \sum_{m=1}^{n-1} f_{2m} + f_{2n}).$$

As with the trapezoidal rule above, the total error is estimated by

$$|I - I_{\text{Simp}}| = \frac{1}{180} (b - a) h^4 |f^{(4)}| + O(h^5),$$

where  $|f^{(4)}|$  represents the upper bound for  $|f^{(4)}(x)|$  over the interval of integration.

<sup>6</sup>This way of approximating the integral is attributed to the English mathematician and inventor Thomas Simpson (1710-1761).

**Solution.** (a) Integration by parts gives the result, i.e.,

$$\begin{aligned} \Gamma(n+1) &= \int_0^\infty e^{-x} x^n dx \\ &= [-x^n e^{-x}]_0^\infty + n \int_0^\infty e^{-x} x^{n-1} dx \\ &= -\lim_{x \rightarrow \infty} \frac{x^n}{e^x} + n\Gamma(n) = 0 + n\Gamma(n) = n\Gamma(n). \end{aligned}$$

Above we used the fact  $\lim_{x \rightarrow \infty} \frac{x^n}{e^x} = 0$ . This limit has the intermediate form  $\frac{\infty}{\infty}$  and a way to compute it occurs by applying the l'Hopital's rule  $n$  times:

$$\lim_{x \rightarrow +\infty} \frac{x^n}{e^x} \stackrel{\infty}{=} \lim_{x \rightarrow +\infty} \frac{n x^{n-1}}{e^x} \stackrel{\infty}{=} \dots \stackrel{\infty}{=} \lim_{x \rightarrow \infty} \frac{n!}{e^x} = 0.$$

(b) Obviously,

$$\Gamma(1) = \int_0^\infty e^{-x} dx = [-e^{-x}]_0^\infty = 1.$$

Moreover, a direct computation shows that

$$\Gamma(2) = \int_0^\infty e^{-x} x dx = 1 = \Gamma(1).$$

Thus we can write  $\Gamma(2) = 1 \cdot \Gamma(1) = 1!$ , and similarly by integration by parts one can prove that  $\Gamma(3) = 2\Gamma(2) = 2!$ . Let us apply induction over  $n$  to prove the general formula. For the inductive step suppose  $\Gamma(n) = (n-1)!$  for some  $n$ . Then, by (a) we see that  $\Gamma(n+1) = n\Gamma(n) = n(n-1)! = n!$ , for all naturals  $n$ .

(c) For any positive integer we have  $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$ . For  $a > 0$  set  $x = ay$  with  $dx = a dy$ . Then we see that

$$\Gamma(n) = \int_0^\infty e^{-ay} (ay)^{n-1} a dy = a^n \int_0^\infty e^{-ay} y^{n-1} dy,$$

i.e.,  $\int_0^\infty e^{-ay} y^{n-1} dy = \frac{\Gamma(n)}{a^n}$ . The result now follows.

(d) This is based on the substitution  $x = u^2$ :

$$\begin{aligned} \Gamma(1/2) &= \int_0^\infty e^{-x} x^{-\frac{1}{2}} dx = \int_0^\infty \frac{e^{-x}}{\sqrt{x}} dx \\ &\stackrel{x=u^2}{dx=2u du} 2 \int_0^\infty e^{-u^2} du = 2 \frac{\sqrt{\pi}}{2} = \sqrt{\pi}. \end{aligned}$$

□

*Recall that the mean value of a function over a given interval provides a single value that represents the average behavior of the function on that interval. This concept is fundamental in various areas of mathematics, since it allows the estimation of the overall effect of a function over an interval without being necessary to consider every individual point. This gives many applications in physics and engineering, where average values are often more useful than point values (such as in calculating average velocity, average temperature, or average concentration of a substance).*



### 3. Sequences, series and limit processes

While building a menagerie of functions, we encountered power series, which extended the collection of all polynomials in a natural way, see 5.4.10. We obtained the class of analytic functions in this way. We have also got the message that they are always smooth and may be integrated and differentiated term by term on the entire domain of convergence. We gave the link to a straightforward technical proof of these statements in 9.4.2 on page 873. Now we shall develop simple tools and show similar results for much more general series of function.

Moreover, functions often depend on further parameters which are dummy when differentiating or integrating, but we need to understand how the result behaves with respect to these parameters. For instance, what about the differentiability of the Gamma function introduced above? Or, when computing volume or area depending on a free parameter, how to minimize it?

Finally, in the end of this chapter we briefly introduce some more advanced concepts of integration.

#### 6.3.1. How well behaved is a sequence of functions? We

return to the discussion of the limits of sequences of functions and the sum of series of functions in view of the methods of differential and integral calculus. Consider a convergent series of functions

$$S(x) = \sum_{n=1}^{\infty} f_n(x)$$

on an interval  $[a, b]$ . Natural questions include:

- If all functions  $f_n(x)$  are continuous at some point  $x_0 \in [a, b]$ , is the function  $S(x)$  also continuous at the point  $x_0$ ?
- If all functions  $f_n(x)$  are differentiable at some point  $a \in [a, b]$ , is the function  $S(x)$  also differentiable there and does the equality  $S'(x) = \sum_{n=1}^{\infty} f'_n(x)$  hold?
- If all functions  $f_n(x)$  are Riemann integrable on an interval  $[a, b]$ , is the function  $S(x)$  also integrable there and does the equality  $\int_a^b S(x) dx = \sum_{n=1}^{\infty} \int_a^b f_n(x) dx$  hold?

Notice, it does not matter whether we discuss series or sequences, since the former ones are just limits of the sequences of the partial sums. First, we demonstrate by examples that the answers to all three questions above are "NO!". Then we find additional conditions on the convergence of the series (or sequences) which guarantees the validity of all three statements. Later we shall mention alternative concepts of integration which are more satisfactory than the Riemann integral for an even wider classes of functions.

#### 6.3.2. Examples of nasty sequences. (1) Consider the functions

$$f_n(x) = (\sin x)^n$$

on the interval  $[0, \pi]$ . The values of these functions are non-negative and smaller than one at all points  $0 \leq x \leq \pi$ , except

On the other hand, the integral mean value theorem provides a rigorous way to calculate the average value of a function over an interval. Next we will describe applications related to these theorems, see also see 6.2.15.

**6.B.54.** For the function  $f(\theta) = \cos(2\theta)e^{1+\sin(2\theta)}$ , determine its average value on  $[-\pi/4, \pi/4]$ .

**Solution.** The average value of  $f$  on  $[-\pi/4, \pi/4]$  is given by

$$m(f) = \frac{1}{\pi/4 - (-\pi/4)} \int_{-\pi/4}^{\pi/4} f(\theta) d\theta = \frac{2}{\pi} \int_{-\pi/4}^{\pi/4} f(\theta) d\theta.$$

To compute this integral one may set  $u = 1 + \sin(2\theta)$ , such that  $du = 2 \cos(2\theta)d\theta$ . We also have  $u = 2$  for  $\theta = \pi/4$ , and  $u = 0$  for  $\theta = -\pi/4$ , thus we get

$$m(f) = \frac{2}{\pi} \cdot \frac{1}{2} \int_0^2 e^u du = \frac{e^2 - 1}{\pi}.$$

Here is a confirmation in Sage:

```
var("th"); f(th)=cos(2*th)*e^(1+sin(2*th))
show((2/pi)*integral(f(th), th, -pi/4, pi/4))
```

□

**6.B.55.** Determine the average velocity  $m(v)$  of a solid in the time interval  $[1, 2]$ , if its velocity is given by

$$v(t) = \frac{t}{\sqrt{1+t^2}}, \quad t \in [1, 2].$$

You can omit the units.

**Solution.** To solve the problem, it suffices to realize that the sought average velocity is the mean value of function  $v$  on interval  $[1, 2]$ . Hence

$$m(v) = \frac{1}{2-1} \int_1^2 \frac{t}{\sqrt{1+t^2}} dt = \int_2^5 \frac{1}{2\sqrt{x}} dx = \sqrt{5} - \sqrt{2}$$

with  $1 + t^2 = x, t dt = dx/2$ . □

Further exercises on improper integrals and other applications related to definite integrals are presented in Section D. Next we describe a few tasks about lengths of curves, areas of regions, and volumes.

**6.B.56. Arc length.** Determine the length of the parametric curve defined by  $x(t) = e^t - t, y(t) = 4e^{t/2}$  with  $0 \leq t \leq 1$ . Then present the solution via Sage.

**Solution.** According to 6.2.18, the length of a curve  $\alpha(t) = [x(t), y(t)]$  on  $\mathbb{R}^2$  defined on the interval  $[a, b]$  is given by the integral

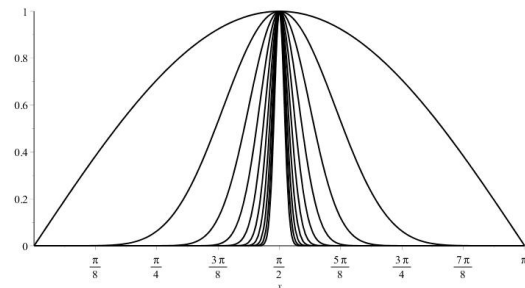
$$s_{\alpha(t)} \equiv s(\alpha(t)) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2} dt.$$

for  $x = \frac{\pi}{2}$ , where the value is 1. Hence on the whole interval  $[0, \pi]$ , these functions converge to the function

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) = \begin{cases} 0 & \text{for all } x \neq \frac{\pi}{2} \\ 1 & \text{for } x = \frac{\pi}{2}. \end{cases}$$

point by point. The limit of the sequence of functions  $f_n$  is a discontinuous function, even though all functions  $f_n(x)$  are continuous. The problematic point is an inner point of the interval.

The same phenomenon occurs for a series of functions, because the sum is the limit of partial sums. Hence in the previous example, it suffices to express  $f_n$  as the  $n$ -th partial sum. For example,  $f_1(x) = \sin x, f_2(x) = (\sin x)^2 - \sin x$ , etc. The figure plots the functions  $f_m(x)$  for  $m = n^3, n = 1, \dots, 10$ .

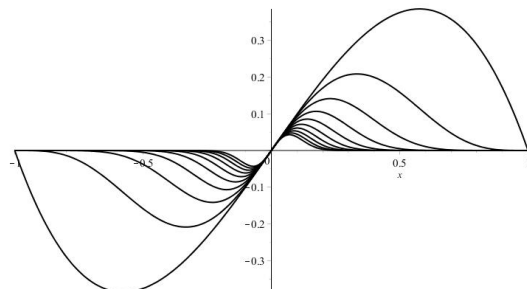


(2) We look at the second question, i.e. badly behaving derivatives. A natural idea based on the same principle as above is to construct a sequence of functions which has the same nonzero derivative at one point, but becomes smaller and smaller. So they converge pointwise to the identically zero function.

The next figure plots the functions

$$f_n(x) = x(1-x^2)^n$$

on the interval  $[-1, 1]$  for values  $n = m^2, m = 1, \dots, 10$ .



It is immediate that  $\lim_{n \rightarrow \infty} f_n(x) = 0$  and that all functions  $f_n(x)$  are smooth. Their derivative at  $x = 0$  is

$$f'_n(0) = ((1-x^2)^n - 2nx^2(1-x^2)^{n-1})|_{x=0} = 1$$

for all  $n$ . But the limit function for the sequence  $f_n$  has a zero derivative at every point.

(3) The counterexample to the third statement is in 6.2.18 already. The characteristic function  $\chi_{\mathbb{Q}}$  of rational numbers can be expressed as a sum of countably many functions, which are numbered exactly by rational numbers. They are zero everywhere except for the single point after which they

For the given curve we have  $x'(t) = e^t - 1$  and  $y'(t) = 2e^{t/2}$ , with  $(x'(t))^2 + (y'(t))^2 = (e^t + 1)^2$ . Thus

$$s = \int_0^1 \sqrt{(e^t + 1)^2} dt = \int_0^1 (e^t + 1) dt = [e^t]_0^1 + [t]_0^1 = e - 1 + 1 = e.$$

In Sage you can confirm all these computations by the following cell:

```
var('t'); x(t)=e^t-t; y(t)=4*e^(t/2)
X(t)=diff(x(t), t); show(X(t))
Y(t)=diff(y(t), t); show(Y(t))
l(t)=(X(t)*X(t)+Y(t)*Y(t)).factor(); show(l(t))
s=integral(sqrt(l(t)), t, 0, 1).simplify_full()
show(s)
```

Using the command `parametric_plot` you can plot the given curve, which can be done just by adding in the previous block the syntax

```
p=parametric_plot((x(t), y(t)), (t, 0, 1))
p.show(aspect_ratio=1/4)
```

□

*It is often useful to jump via Sage from classical integration to numerical integration. This for example can be useful for cases where we need an approximation, or in cases where a “closed form” of the integral does not exist, as for example the case of Gaussian  $f(x) = e^{-x^2}$  whose antiderivative is the so-called “error function”  $\operatorname{erf}$ . This is an interesting situation which, till the end of this chapter, we will implement in many different ways.*

*Notice in Sage an option to approximate a definite integral relies on the command `n(integral(f, x, a, b))`, or `numerical_approx(integral(f, x, a, b))`. For instance for the Gaussian  $f(x) = e^{-x^2}$  try the cell*

```
f(x)=e^(-x^2)
show(integral(f, x, 0, 1))
print(n(integral(f, x, 0, 1)))
```

Sage’s output has the form

$$\frac{1}{2} \sqrt{\pi} \operatorname{erf}(1)$$

and 0.746824132812427, respectively. Built-in methods that Sage provides for numerical integration will be discussed later. We begin with the following task about the arc length of parametric curves.

**6.B.57.** Given a curve  $c : [a, b] \rightarrow \mathbb{R}^2$  in the parametric form  $c(t) = [x(t), y(t)]$ , present a routine in Sage that will compute numerically the length of  $c$  on  $[a, b]$  and plot its graph. Then, test your routine for the following cases and next confirm routine’s result by a formal computation:

- (1)  $c(t) = [\cos^3(t), \sin^3(t)]$ ,  $t \in [0, \pi/2]$ ;
- (2)  $c(t) = [t, \ln(\cos(t))]$ ,  $t \in [0, \pi/4]$ ;

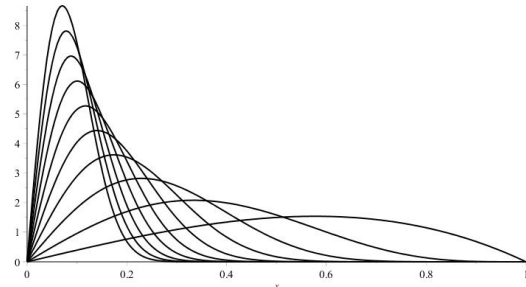
are named for, where the value is 1. Riemann integrals of all such functions are zero, but the sum is not a Riemann integrable function.

This example illustrates the fundamental flaw of the Riemann integral, to which we return later.

We present an example where the limit function  $f$  is integrable, all functions  $f_n$  are continuous, but the value of the integral is not the limit of the integrals of  $f_n$ . We modify the sequence of the functions  $x(1 - x^2)^n$  used above. They integrate to  $\int_0^1 x(1 - x^2)^n dx = \frac{1}{2(n+1)}$ . Thus, we consider the functions

$$f_n(x) = 2(n+1)x(1 - x^2)^n.$$

These functions with  $n = m^2$ ,  $m = 1, \dots, 10$  are on the next diagram.



It is quite easy to verify that the values of these functions converge to zero for every  $x \in [0, 1]$  (for example the reader can check  $\ln(f_n(x)) \rightarrow -\infty$ ). But for all  $n$

$$\int_0^1 f_n(x) dx = 1 \neq 0.$$

**6.3.3. Uniform convergence.** A reason of failure in all three previous examples is the fact that the speed of pointwise convergence of values  $f_n(x) \rightarrow f(x)$  varies dramatically point from point. Hence a natural idea is to confine the problem to cases where the convergence will have roughly the same speed over all the interval:



#### UNIFORM CONVERGENCE

**Definition.** We say that the sequence of functions  $f_n(x)$  converges uniformly on interval  $[a, b]$  to the limit  $f(x)$ , if for every positive number  $\varepsilon$ , there exists a natural number  $N \in \mathbb{N}$  such that for all  $n \geq N$  and all  $x \in [a, b]$  the inequality

$$|f_n(x) - f(x)| < \varepsilon$$

holds.

A series of functions converges uniformly on an interval, if the sequence of its partial sums converges uniformly.

Albeit the choice of the number  $N$  depends on the chosen  $\varepsilon$ , it is independent on the point  $x \in [a, b]$ . This is the difference from pointwise convergence, where  $N$  depends on both  $\varepsilon$  and  $x$ . We visualise the definition graphically in this way: if we consider a zone created by a translation of the limit function  $f(x)$  to  $f(x) \pm \varepsilon$  for arbitrarily small, but fixed positive

- (3)  $c(t) = [t \sin(t), t \cos(t)]$ ,  $t \in [0, 4\pi]$ ;  
 (4)  $c(t) = [\sin^2(t), \cos^2(t)]$ ,  $t \in [0, \frac{\pi}{2}]$ . ○

**6.B.58.** Compute the length  $s$  of a part of the so-called “trac-trix”, that is, the parametric curve  $\alpha(t) = [f(t), g(t)]$  defined by

$$f(t) = r \cos t + r \ln \left( \tan \left( \frac{t}{2} \right) \right), \quad g(t) = r \sin(t),$$

with  $t \in [\pi/2, a]$ , where  $r > 0$ , and  $a \in (\pi/2, \pi)$ . Next use Sage to plot  $\alpha$  for the values  $r = 1, 2, \dots, 5$  in one figure. ○

**6.B.59. Area of half ellipses.** Consider the curve  $\alpha(t) = [x(t), y(t)] = [4 \cos(t), 3 \sin(t)]$ , with  $0 \leq t \leq \pi$ .

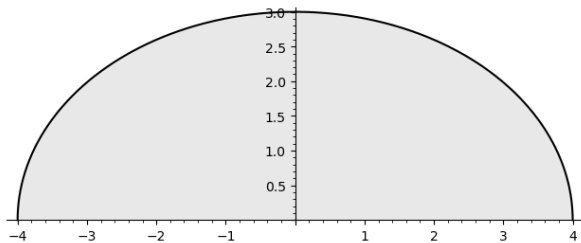
(a) Show that  $\alpha(t)$  is the half of a horizontal ellipses centered at the origin, with major axes of length 8 and minor axes of length 6. Use Sage to plot  $\alpha$  and color the region bounded by the  $x$ -axis and the ellipsis.

(b) Compute the area of this region.

**Solution.** (a) Obviously,  $\alpha(t)$  is the half of a horizontal ellipses centered at the origin of  $\mathbb{R}^2$ , since it satisfies the equation  $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$  with  $a = 4$  and  $b = 3$ . Thus the major axes has length  $2a = 8$  and the minor axes has length  $2b = 6$ , respectively. To plot  $\alpha$  we will use the command `parametric_plot` as above, that is,

```
x(t)=4*cos(t); y(t)=3*sin(t)
show(parametric_plot((x(t), y(t)), (t, 0, pi),
fill=True, fillcolor="lightgrey",
aspect_ratio=1, color="black", thickness=1.5))
```

which produces the following figure:



(b) To find the area under the given parametric curve we need first to compute the integral

$$\begin{aligned} \int_0^\pi y(t)x'(t) dt &= \int_0^\pi 3 \sin(t)(4 \cos(t))' dt \\ &= -12 \int_0^\pi \sin^2(t) dt = -\frac{12}{2} \int_0^\pi (1 - \cos(2t)) dt \\ &= -6 \left[ t - \frac{\sin(2t)}{2} \right]_0^\pi = -6\pi. \end{aligned}$$

Thus the required area equals to  $|\int_0^\pi y(t)x'(t) dt| = 6\pi$ . In Sage to confirm this result it is sufficient to add in the previous block the cell

```
F(t)=y(t)*diff(x(t), t)
show(abs(integral(F(t), t, 0, pi)))
```

$\varepsilon$ , all of the functions  $f_n(x)$  will fall into this zone, except for finitely many of them. The first and the last of the nasty examples above do not have this property; In the second example, the sequence of derivatives  $f'_n$  lacked it.

**6.3.4.** The three claims in the following theorem say that all three generally false properties discussed in 6.3.1 are true features for uniform convergence (but beware the subtleties when differentiating).



CONSEQUENCES OF UNIFORM CONVERGENCE

**Theorem.** (1) Let  $f_n(x)$  be a sequence of functions continuous on a closed interval  $[a, b]$  and converging uniformly to the function  $f(x)$  on this interval. Then  $f(x)$  is also continuous on the interval  $[a, b]$ .

(2) Let  $f_n(x)$  be a sequence of Riemann integrable functions on a finite closed interval  $[a, b]$  which converge uniformly to the function  $f(x)$  on this interval. Then  $f(x)$  is Riemann integrable, and

$$\int_a^b f(x) dx = \int_a^b \left( \lim_{n \rightarrow \infty} f_n(x) \right) dx = \lim_{n \rightarrow \infty} \int_a^b f_n(x) dx.$$

(3) Let  $f_n(x)$  be a sequence of functions differentiable on a closed interval  $[a, b]$  and assume  $f_n(x_0) \rightarrow f(x_0)$  at some point  $x_0 \in [a, b]$ . Moreover, assume all derivatives  $g_n(x) = f'_n(x)$  are continuous and converge uniformly to the function  $g(x)$  on the same interval. Then the function  $f(x) = f(x_0) + \int_{x_0}^x g(t) dt$  is differentiable on the interval  $[a, b]$ , the functions  $f_n(x)$  converge to  $f(x)$  and  $f'(x) = g(x)$ . In other words,

$$\frac{d}{dx} f(x) = \frac{d}{dx} \left( \lim_{n \rightarrow \infty} f_n(x) \right) = \lim_{n \rightarrow \infty} \left( \frac{d}{dx} f_n(x) \right)$$

**PROOF OF THE FIRST CLAIM.** Fix an arbitrary fixed point  $x_0 \in [a, b]$  and let  $\varepsilon > 0$  be given. It is required to show that

$$|f(x) - f(x_0)| < \varepsilon$$

for all  $x$  close enough to  $x_0$ . From the definition of uniform convergence,

$$|f_n(x) - f(x)| < \varepsilon$$

for all  $x \in [a, b]$  and all sufficiently large  $n$ . Choose some  $n$  with this property and consider  $\delta > 0$  such that

$$|f_n(x) - f_n(x_0)| < \varepsilon$$

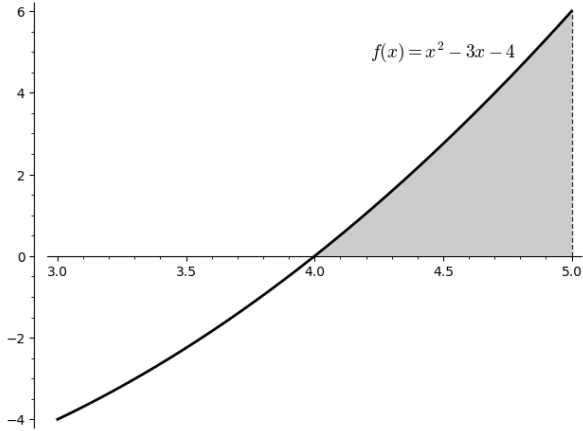
for all  $x$  in  $\delta$ -neighbourhood of  $x_0$ . That is possible because  $f_n(x)$  are continuous for all  $n$ . Then

$$\begin{aligned} |f(x) - f(x_0)| &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(x_0)| \\ &\quad + |f_n(x_0) - f(x_0)| < 3\varepsilon \end{aligned}$$

for all  $x$  in the  $\delta$ -neighbourhood of  $x_0$ . This is the desired inequality with the bound  $3\varepsilon$ . □

**6.B.60. Area.** Calculate the area  $E$  between the  $x$ -axis and the graph of  $f(x) = x^2 - 3x - 4$  for  $4 \leq x \leq 5$ . Next use Sage to illustrate the area in question.

**Solution.** An easy computation shows that  $E = \frac{17}{6}$ , and this corresponds to grey region in the figure given here:



To obtain the required illustration in Sage, we will use the polygon method, which is the core of our program presented below. Details on this method are postponed to Section D, see (b) in 6.D.43.

```
f(x)=x^2-3*x-4
p=plot(f(x), x, 3, 5, thickness=2,
color="black")
p+=line([(5, 0), (5, 6)],
rgbcolor=(0.2,0.2,0.2), linestyle="--")
p+=polygon([(4, 0), (4, f(4))]
+ [(x, f(x)) for x in [4, 4.1, ..., 5]]+[(5, 0), (4, 0)])
rgbcolor=(0.8,0.8,0.8), aspect_ratio='automatic')
p+=text("$f(x)=x^2-3x-4$",
(4.5, 5), fontsize=14, color="black")
show(p)
show(integral(f(x), x, 4, 5))
```

**6.B.61.** Compute the volume of a solid created by rotation of a bounded surface, whose boundary is the curve  $x^4 - 9x^2 + y^4 = 0$ , around the  $x$ -axis. You may omit the units.

**Solution.** If  $[x, y]$  is a point on the  $x^4 - 9x^2 + y^4 = 0$ , clearly this curve also intersects points  $[-x, y]$ ,  $[x, -y]$ ,  $[-x, -y]$ . Thus is symmetric with respect to both axes  $x, y$ . For  $y = 0$ , we have  $x^2(x - 3)(x + 3) = 0$ , i.e. the  $x$  axis is intersected by the boundary curve at points  $[-3, 0]$ ,  $[0, 0]$ ,  $[3, 0]$ . In the first quadrant, it can then be expressed as a graph of the function

$$f(x) = \sqrt[4]{9x^2 - x^4}, \quad x \in [0, 3].$$

The sought volume is thus a double (here we consider  $x > 0$ ) of the integral

$$\int_0^3 \pi f^2(x) dx = \pi \int_0^3 \sqrt{9x^2 - x^4} dx.$$

**Remark.** In fact, the arguments in the proof show a more general claim. Indeed, if the functions  $f_n(x)$  converge uniformly to  $f(x)$  on  $[a, b]$ , and the individual functions  $f_n(x)$  have the limits (or one-sided limits)  $\lim_{x \rightarrow x_0} f_n(x) = a_n$ , then the limit  $\lim_{x \rightarrow x_0} f(x)$  exists if and only if the limit  $\lim_{n \rightarrow \infty} a_n = a$  exists. Then they are equal, that is,

$$a = \lim_{n \rightarrow \infty} \left( \lim_{x \rightarrow x_0} f_n(x) \right) = \lim_{x \rightarrow x_0} \left( \lim_{n \rightarrow \infty} f_n(x) \right).$$

The reader should be able to modify the above proof for this situation.

**6.3.5. Proof of the second claim.** The proof of this part of the theorem is based upon a generalization of the properties of Cauchy sequences of numbers to uniform convergence of functions. In this way we can work with the existence of the limit of a sequence of integrals without needing to know the limit.



UNIFORMLY CAUCHY SEQUENCES

**Definition.** The sequence of functions  $f_n(x)$  on interval  $[a, b]$  is *uniformly Cauchy*, if for every (small) positive number  $\varepsilon$ , there exists a (large) natural number  $N$  such that for all  $x \in [a, b]$  and all  $m, n \geq N$ ,

$$|f_n(x) - f_m(x)| < \varepsilon.$$

Every uniformly convergent sequence of functions on interval  $[a, b]$  is also uniformly Cauchy on the same interval. To see this, it suffices to notice the usual bound

$$|f_n(x) - f_m(x)| \leq |f_n(x) - f(x)| + |f(x) - f_m(x)|$$

based on the triangle inequality.

Before coming to the proof 6.3.4(2), we mention the following:

**Proposition.** Every uniformly Cauchy sequence of functions  $f_n(x)$  on the interval  $[a, b]$  uniformly converges to some function  $f$  on this interval. □

**PROOF.** Of course, the condition for a sequence of functions to be uniformly Cauchy implies that also for all  $x \in [a, b]$ , the sequence of values  $f_n(x)$  is a Cauchy sequence of real (or complex) numbers. Hence the sequence of functions  $f_n(x)$  converges pointwise to some function  $f(x)$ .

Choose  $N$  large enough so that

$$|f_n(x) - f_m(x)| < \varepsilon$$

for some small positive  $\varepsilon$  chosen beforehand and all  $m, n \geq N, x \in [a, b]$ . Now choose one such  $n$  and fix it, then

$$|f_n(x) - f(x)| = \lim_{m \rightarrow \infty} |f_n(x) - f_m(x)| \leq \varepsilon$$

for all  $x \in [a, b]$ . Hence the sequence  $f_n(x)$  converges to its limit uniformly. □

**PROOF OF THE SECOND CLAIM IN 6.3.4.** Recall that every uniformly convergent sequence of functions is also uniformly Cauchy and that the Riemann sums of all single terms  $f_n(x)$

Using the substitution  $t = \sqrt{9 - x^2}$  ( $x dx = -t dt$ ), we get

$$\int_0^3 \sqrt{9x^2 - x^4} dx = \int_0^3 x \cdot \sqrt{9 - x^2} dx = - \int_3^0 t^2 dt = 9.$$

Thus the final answer is  $18\pi$ . □

**6.B.62. Torricelli's trumpet (1641).** Let a part of a branch of the hyperbola  $xy = 1$  for  $x \geq a$ , where  $a > 0$ , rotate around the  $x$  axis. Show that the solid of revolution created in this manner has a finite volume  $V$  and simultaneously an infinite surface  $S$ .

**Solution.** We know that

$$\begin{aligned} V &= \pi \int_a^{+\infty} \left(\frac{1}{x}\right)^2 dx = \pi \int_a^{+\infty} \frac{1}{x^2} dx \\ &= \pi \left( \lim_{x \rightarrow +\infty} -\frac{1}{x} - \left(-\frac{1}{a}\right) \right) = \frac{\pi}{a} \end{aligned}$$

and

$$\begin{aligned} S &= 2\pi \int_a^{+\infty} \frac{1}{x} \cdot \sqrt{1 + \left(-\frac{1}{x^2}\right)^2} dx = 2\pi \int_a^{+\infty} \frac{\sqrt{x^4 + 1}}{x^3} dx \\ &\geq 2\pi \int_a^{+\infty} \frac{1}{x} dx = 2\pi \left( \lim_{x \rightarrow +\infty} \ln x - \ln a \right) = +\infty. \end{aligned}$$

The fact that the given solid (the so called Torricelli's trumpet) cannot be painted with a finite amount of colour, but can be filled with a finite amount of fluid, is called the "*Torricelli's paradox*". Realize however that a real color painting has a nonzero width, which the computation does not take into account. For example, if we would paint it from the inside, a single drop of color would undoubtedly "block" the trumpet of infinite length. □

*A differential equation is an equation involving derivatives. Many physical and engineering problems naturally lead to differential equations, making it essential to learn how to formulate and solve them. For instance, in Chapter 3 we encountered differential equations in the context of population growth.*

Differential equations fall into two main categories, "ordinary differential equations" (ODEs) and "partial differential equations" (PDEs). An ODE involves derivatives with respect to a single independent variable, and next we focus on ODEs of first order. Typically, such equations are written as  $y' = F(x, y)$ , where  $y = f(x)$ , see 6.2.21. If  $F$  is in the form  $y' = F(x, y) = f(x)g(y)$  for some functions  $f, g$ , the ODE is called separable. For  $g(y) \neq 0$  we can solve it by separating the variables:

$$\frac{dy}{g(y)} = f(x) dx \Rightarrow \int \frac{dy}{g(y)} = \int f(x) dx.$$

Let's explore some simple problems, with a more detailed discussion on differential equations postponed to Chapter 8.

of the sequence converge to  $\int_a^b f_n(x) dx$  independently of the choice of the partition and the representatives. Hence, if

$$|f_n(x) - f_m(x)| < \varepsilon$$

for all  $x \in [a, b]$ , then also

$$\left| \int_a^b f_n(x) dx - \int_a^b f_m(x) dx \right| \leq \varepsilon |b - a|.$$

Therefore the sequence of numbers  $\int_a^b f_n(x) dx$  is Cauchy, and hence convergent.

The Riemann sums of the limit function  $f(x)$  can be made arbitrarily close to those of  $f_n(x)$  for large  $n$ , by the same argument as above. So  $f(x)$  is integrable. Moreover,

$$\left| \int_a^b f_n(x) dx - \int_a^b f(x) dx \right| \leq \varepsilon |b - a|,$$

so the limit value is as expected. □

**6.3.6. Proof of the third claim.** For the corresponding result about derivatives, extra care is needed regarding the assumptions: If the differentiable functions  $\tilde{f}_n(x) = f_n(x) - f_n(x_0)$  are considered instead of  $f_n(x)$ , the derivatives do not change. Hence without loss of generality it can be assumed that all functions satisfy  $f_n(x_0) = 0$ .

Then the first assumption of the theorem is satisfied automatically. For all  $x \in [a, b]$ , we can write

$$f_n(x) = \int_{x_0}^x g_n(t) dt.$$

Because the functions  $g_n$  converge uniformly to  $g$  on all of  $[a, b]$ , the functions  $f_n(x)$  converge to

$$f(x) = \int_{x_0}^x g(t) dt.$$

$g$  is a uniform limit of continuous functions, thus  $g$  is again continuous. By 6.2.8, for the relations between the Riemann integral and the primitive function, the proof is finished.

**6.3.7. Uniform convergence of series.** For infinite series, we apply the previous three results to the sequences of partial sums. Thus the following corollary is an immediate consequence:



**6.B.63. Differential equations.** Compute the limit  $\lim_{x \rightarrow +\infty} f(x)$  if it is known that the function  $f(x)$  satisfies the differential equation

$$f'(x) + \alpha \cdot e^x (f(x))^2 = 0,$$

with initial condition  $f(0) = 1/\alpha$ , where  $\alpha > 0$  is some constant. Next use Sage to solve the given differential equation for  $\alpha = 3$ .

**Solution.** Let us write  $y = f(x)$ . Then, the given differential equation takes the form as  $y' = -\alpha \cdot e^x y^2$ , or equivalently

$$\frac{dy}{dx} = -\alpha \cdot e^x y^2 \iff -\frac{dy}{y^2} = \alpha \cdot e^x dx.$$

Hence, by integrating we obtain

$$-\int \frac{dy}{y^2} = \alpha \int e^x dx,$$

which is equivalent to

$$-\int y^{-2} dy = \alpha e^x + C$$

that is,  $y^{-1} = \alpha e^x + C$ . Thus  $y = f(x) = \frac{1}{\alpha e^x + C}$  and to find  $C$  one relies on the initial condition  $f(0) = 1/\alpha$ , which gives  $C = 0$ . Thus  $f(x) = \frac{1}{\alpha e^x}$  and we see that  $\lim_{x \rightarrow +\infty} f(x) = 0$ .

Recall by 6.B.5 that a method to solve (1st order) ODEs relies on the command `desolve`. Since  $a = 3$  in this case one can type

```
var("x"); y = function("y")(x)
a=3; show(desolve(diff(y, x) + a*e^x*y^2, y))
```

and Sage's output has the form  $\frac{1}{3y(x)} = C + e^x$ , which is of course equivalent to our answer for  $a = 3$ .  $\square$

**6.B.64.** Solve the ODE  $\frac{dy}{dx} = e^{\sin(x)} \cos(x) y^{1/2}$  with initial condition  $y(0) = 16$ .  $\circ$

**6.B.65.** Solve the task in 6.B.64 using Sage. Moreover, plot the solution together with the "slope field" of the corresponding differential equation.

**Solution.** Recall by 6.B.5 that in order to solve an initial value problem in Sage we use the `desolve` command and include the option `ics = [x0, y0]` (or simply `[x0, y0]`), where  $y_0 = y(x_0)$  represents the initial condition. Thus, in our case one can type

```
var("x")
y = function("y")(x)
h=desolve(diff(y, x) - e^(sin(x))*cos(x)*sqrt(y),
y, ics=[0, 16])
show(h)
```

which prints out the expression  $2\sqrt{y(x)} = e^{\sin(x)} + 7$ , hence confirms our result in 6.B.64.

Having a differential equation of the form  $y' = F(x, y)$ , plotting the slope field means that we plot a line at  $(x, y)$  with slope  $F(x, y)$ . Such details will be analyzed more carefully in Chapter 8. However, we may already discuss the built-in method that Sage provides for this procedure.

CONSEQUENCES FOR UNIFORM CONVERGENCE OF SERIES

**Theorem.** Consider a sequence of functions  $f_n(x)$  on interval  $[a, b]$ .

(1) If all the functions  $f_n(x)$  are continuous on  $[a, b]$  and the series

$$S(x) = \sum_{n=1}^{\infty} f_n(x)$$

converges uniformly to the function  $S(x)$ , then  $S(x)$  is continuous on  $[a, b]$ .

(2) If all the functions  $f_n(x)$  are Riemann integrable on  $[a, b]$  and the series

$$S(x) = \sum_{n=1}^{\infty} f_n(x)$$

uniformly converges to  $S(x)$  on  $[a, b]$ , then  $S(x)$  is integrable on  $[a, b]$  and

$$\int_a^b \left( \sum_{n=1}^{\infty} f_n(x) \right) dx = \sum_{n=1}^{\infty} \int_a^b f_n(x) dx.$$

(3) If all the functions  $f_n(x)$  are continuously differentiable on the interval  $[a, b]$ , if the series  $S(x) = \sum_{n=1}^{\infty} f_n(x)$  converges for some  $x_0 \in [a, b]$ , and if the series  $T(x) = \sum_{n=1}^{\infty} f'_n(x)$  converges uniformly on  $[a, b]$ , then the series  $S(x)$  converges.  $S(x)$  is continuously differentiable on  $[a, b]$  and  $S'(x) = T(x)$ . That is:

$$\frac{d}{dx} \left( \sum_{n=1}^{\infty} f_n(x) \right) = \sum_{n=1}^{\infty} \frac{d}{dx} f_n(x).$$

**6.3.8. Test of uniform convergence.** A simple way to test that a series of functions converges uniformly is to use a comparison with the absolute convergence of a suitable series of numbers. This is often called the *Weierstrass test*.

Suppose a sequence of functions  $f_n(x)$  is given on an interval  $I = [a, b]$  satisfying

$$|f_n(x)| \leq a_n \in \mathbb{R}$$

for suitable real constants  $a_n$  and for all  $x \in [a, b]$ . Let

$$s_k(x) = \sum_{n=1}^k f_n(x)$$

for distinct indices  $k$ . For  $k > m$ ,

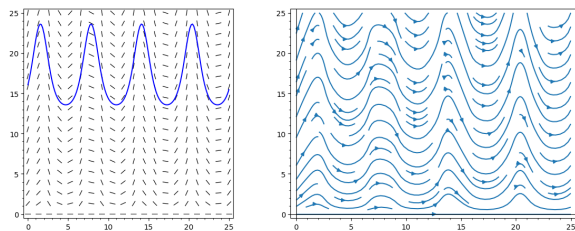
$$\begin{aligned} |s_k(x) - s_m(x)| &= \left| \sum_{n=m+1}^k f_n(x) \right| \\ &\leq \sum_{n=m+1}^k |f_n(x)| \leq \sum_{n=m+1}^k a_n. \end{aligned}$$



The latter is based on the syntax `plot_slope_field`, but first we need to introduce the function  $F$  representing the differential equation. One can also plot the solution curves (“integral curves”), as slope fields, via the command `streamline_plot(F(x, y), (x, a, b), (y, a, b))`. For our task, we have  $\frac{dy}{dx} = e^{\sin(x)} \cos(x)y^{1/2}$ , thus we need to consider the function  $F(x, y) = e^{\sin(x)} \cos(x)y^{1/2}$ , and the implementation takes the form

```
var("x, y")
c1=implicit_plot(2*sqrt(y)==e^(sin(x))+7,
(x, 0, 25), (y, 0, 25))
F(x, y)=e^(sin(x))*cos(x)*sqrt(y)
p1=plot_slope_field(F(x, y), (x, 0, 25), (y, 0, 25))
p2=streamline_plot(F(x, y), (x, 0, 25), (y, 0, 25))
show(p1+c1, figsize=6); show(p2, figsize=6)
```

This produces the following figures



Here, at the left hand side with blue appears the solution of our initial value problem, while the slope field has grey colour and appears in the background. At the right hand side we see the integral curves, as slope fields. □

*Numerical integration is a fundamental technique in computational mathematics, used to approximate the area under a curve, especially when an exact analytical solution is challenging or impossible to obtain. It complements the numerical differentiation that we met earlier in 6.1.12 and is invaluable for analyzing functions where direct solutions are impractical. Below, we will explore examples using the trapezoid rule and Simpson's rule, see 6.2.23.*

**6.B.66. Trapezoid rule.** Compute the error of estimating the integral

$$I = \int_0^\pi \sin(x) dx$$

by the trapezoid rule with  $n = 4$  intervals. Demonstrate this by the following two ways:

- i) Compute the actual error given by the difference  $|I - I_{\text{trap}}|$ .
- ii) Apply the formula  $\frac{(b-a)h^2}{12} |f''|$ , where  $|f''|$  represents an upper estimate of  $|f''(x)|$  over the integral of integration  $[a, b]$ , where  $f(x) = \sin(x)$ .

**Solution.** The trapezoid rule with  $n = 4$  intervals has the form

$$I_{\text{trap}} = \frac{h}{2} [f(x_0) + 2f(x_1) + 2f(x_2) + 2f(x_3) + f(x_4)],$$

If the series of the (nonnegative) constants  $\sum_{n=1}^\infty a_n$  is convergent, then the sequence of its partial sums is a Cauchy sequence. But then the sequence of partial sums  $s_n(x)$  is uniformly Cauchy.

By 6.3.5 the following is verified:

THE WEIERSTRASS TEST

**Theorem.** Let  $f_n(x)$  be a sequence of functions defined on interval  $[a, b]$  with  $|f_n(x)| \leq a_n \in \mathbb{R}$ .

If the series of numbers  $\sum_{n=1}^\infty a_n$  is convergent, then the series  $S(x) = \sum_{n=1}^\infty f_n(x)$  converges uniformly.

**6.3.9. Consequences for power series.** The Weierstrass test allows us to derive the properties of power series in a very straightforward way. Consider a series



$$S(x) = \sum_{n=0}^\infty a_n(x - x_0)^n$$

centered at a point  $x_0$ .

We saw earlier in 5.4.9, that each power series converges on an entire interval  $(x_0 - \delta, x_0 + \delta)$ . The radius of convergence  $\delta \geq 0$  can be zero or  $\infty$ , see 5.4.13. Moreover, the series obtained by integrating or differentiating a power series term by term must have the same radius of convergence (by the very formula based on  $\limsup$ ).

In the proof of theorem 5.4.9, a comparison with a suitable geometric series is used to verify the convergence of power series. By the Weierstrass test, every power series  $S(x)$  converges uniformly on every compact (i.e. bounded closed) interval  $[a, b]$  contained in the interval  $(x_0 - \delta, x_0 + \delta)$ . Thus the crucial result follows again:

DIFFERENTIATION AND INTEGRATION OF POWER SERIES

**Theorem.** Every power series  $S(x)$  is continuous and is continuously differentiable at all points inside its interval of convergence. The function  $S(x)$  is Riemann integrable and can be differentiated or integrated done term by term.

*Abel's theorem* states that power series are continuous even at the boundary points of their domain when they converge there (including eventual infinite limits). We do not prove it here.

The pleasant properties of power series also reveal limitations on the use in practical modelling. In particular, it is not possible to approximate piece-wise continuous or non-differentiable functions very well by using power series. Of course, it should be possible to find better sets of functions  $f_n(x)$  than just the values  $f_n(x) = x^n$ , up to constants. The best known examples are Fourier series and wavelets discussed in the next chapter.

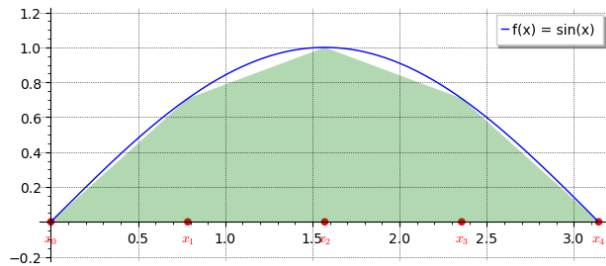
where the step size  $h$  satisfies  $h = \frac{b-a}{n}$ , see 6.2.23. For our case we have  $f(x) = \sin(x)$ ,  $a = 0$ ,  $b = \pi$  and  $h$  equals  $\pi/4$ . Thus,  $x_0 = 0$ ,  $x_1 = \pi/4$ ,  $x_2 = \pi/2$ ,  $x_3 = 3\pi/4$ , and  $x_4 = \pi$ , with  $f(x_0) = 0 = f(x_4)$ ,  $f(x_1) = \frac{\sqrt{2}}{2} = f(x_3)$  and  $f(x_2) = 1$ . This gives

$$I_{\text{trap}} = \frac{\pi}{8} [2(1 + \sqrt{2})] \approx 1.896,$$

where we used Sage to compute this expression by the command

```
N((pi/8)*2*(1+sqrt(2)))
```

An illustration of the trapezoid rule is given here



Let us now derive the actual error of the estimation as suggested in (i). You can easily compute  $I = 2$ , thus

$$|I - I_{\text{trap}}| \approx \left| \int_0^\pi \sin(x) dx - 1.896 \right| \approx 0.104.$$

This error should not differ dramatically from the result in (ii). Indeed, we see that  $f''(x) = -\sin(x)$  and over the interval  $[0, \pi]$  the maximum value of  $|f''(x)| = |\sin(x)|$  equals 1. This gives us

$$\frac{(b-a)^3}{12n^2} \max_{x \in [0, \pi]} |f''(x)| \stackrel{n=4}{=} \frac{\pi^3}{192} \approx 0.162.$$

Therefore, the actual error 0.104 is less than the theoretical upper bound 0.162, which is consistent with the idea that the theoretical error estimate often provides an upper bound on the actual error.

We also conclude that the trapezoid rule underestimates the actual integral value. This occurs because  $f(x) = \sin(x)$  is concave (down) on the interval of integration, i.e.,  $f''(x) < 0$  for all  $x \in [0, \pi]$ . This characteristic is typical of the trapezoid rule's behaviour with concave functions, as illustrated in the figure above. □

**6.B.67.** Create a routine in Sage to demonstrate the computation of the trapezoid rule with  $n$  intervals. Apply your program to estimate the integral

$$L = \int_{\pi/2}^{3\pi/2} \cos(x) dx,$$

with  $n = 4$ . Additionally, provide a formal proof, and calculate both the actual and theoretical errors of the estimation, as discussed in 6.B.66. ○

**6.B.68.** Apply the Simpson's rule to approximate the integral

$$I = \int_0^1 \frac{1}{1+x} dx \text{ with } h = 1/4 \text{ and } h = 1/2 \text{ respectively.}$$

**6.3.10. Laurent series.** We return to the smooth function  $f(x) = e^{-1/x^2}$  from paragraph 6.1.9 in the context of Taylor series expansions. It is not analytic at the origin, because all its derivatives are zero there and the function is strictly positive at all other points. At all points  $x_0 \neq 0$  this function is given by its convergent Taylor series with radius of convergence  $r = |x_0|$ . At the origin the Taylor series converges only at the one point 0.



Replace  $x$  with the expression  $-1/x^2$  into the power series for  $e^x$ . The result is the series of functions

$$S(x) = \sum_{n=0}^{\infty} \frac{1}{n!} (-1)^n x^{-2n} = \sum_{n=-\infty}^0 \frac{(-1)^{|n|}}{|n|!} x^{2n}.$$

The series converges at all points  $x \neq 0$ . It gives a good idea about the behaviour near the exceptional point  $x = 0$ .

Thus we consider the following series similar to power series but more general:

### LAURENT<sup>7</sup> SERIES

A series of functions of the form

$$S(x) = \sum_{n=-\infty}^{\infty} a_n (x - x_0)^n$$

is called a *Laurent series centered at  $x_0$* . The series is convergent if both its parts with positive and negative exponents converge separately.

The importance of Laurent series can be seen with rational functions. Consider such a function  $S(x) = f(x)/g(x)$  with coprime polynomials  $f$  and  $g$  and consider a root  $x_0$  of polynomial  $g(x)$ . If the multiplicity of this root is  $s$ , then after multiplication we obtain the function  $\tilde{S}(x) = S(x)(x - x_0)^s$ , which is analytic on some neighbourhood of  $x_0$ . Therefore we can write

$$\begin{aligned} S(x) &= \frac{a_{-s}}{(x-x_0)^s} + \dots + \frac{a_{-1}}{x-x_0} + a_0 + a_1(x-x_0) + \dots \\ &= \sum_{n=-s}^{\infty} a_n (x-x_0)^n. \end{aligned}$$

Consider the two parts of the Laurent series separately:

$$S(x) = S_- + S_+ = \sum_{n=-\infty}^{-1} a_n (x-x_0)^n + \sum_{n=0}^{\infty} a_n (x-x_0)^n.$$

For the series  $S_+$ , Theorem 5.4.9 implies that its radius of convergence  $R$  is given by  $R^{-1} = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$ . Apply the same idea to the series  $S_-$  with  $1/x$  substituted for  $x$ . It is then apparent that the series  $S_-(x)$  converges for  $|x-x_0| > r$ , where  $r = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_{-n}|}$ .

<sup>7</sup>Pierre Alphonse Laurent (1813-1854) was a French engineer and military officer. He submitted his generalization of the Taylor series into the Grand Prix competition of the French Académie des Sciences. For formal reasons it was not considered. It was published much later, after the author's death.

Then, compare the accuracy of these results with those obtained using the trapezoidal rule. Which method provides a better approximation? ○

**6.B.69.** Create a routine in Sage to demonstrate the computation of the Simpson’s rule with  $n$  intervals, where  $n$  is an even integer. Next use your program to verify the result presented in 6.B.68 for  $h = 1/4$ , and illustrate the Simpson’s approximation. ○

**6.B.70.** Calculate the theoretical error  $\frac{(b-a)^5}{180 n^4} |f^{(4)}|$  of Simpson’s approximation with  $h = 1/4$  for the integral  $I$  given in 6.B.68, where  $|f^{(4)}|$  represents the upper bound of  $|f^{(4)}(x)|$  with  $x \in [0, 1]$ . Then, compare this theoretical error with the actual error computed earlier. ○

**C. Sequences, series and limit processes**

*We are already familiar with sequences of real numbers, power series and the concept of convergence. Our aim now is to readdress the discussion on series of functions in light of the methods of differential and integral calculus.*

*First we will consider sequences whose terms are functions rather than real or complex numbers. Sequences of functions naturally arise in real analysis and are crucial in approximation theory. We will illustrate the consequences of the uniform convergence of sequences and series of functions through numerous examples. Additionally, we will discuss tasks related to the differentiation and integration of power series, along with other applications. Let us begin, however, with numerical series. Thanks to the integral criterion of convergence (see 6.2.15), one can address the question of convergence for a broader class of series. The next few tasks will highlight this fact.*

**6.C.1. Applications of the integral criterion of convergence.** Decide whether the following sums converge or diverge and confirm your computations in Sage:

$$T_1 = \sum_{n=1}^{\infty} \frac{1}{n \ln n}, \quad T_2 = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

**Solution.** Observe that we cannot decide the convergence of none of these series by using the ratio or root test. This is because

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n \rightarrow \infty} \sqrt[n]{a_n} = 1.$$

However, using the integral criterion for convergence of series one obtains:

$$\int_1^{\infty} \frac{1}{x \ln(x)} dx = \int_0^{\infty} \frac{1}{t} dt = \lim_{\delta \rightarrow \infty} [\ln(t)]_0^{\delta} = \infty,$$

hence the series  $S_1$  diverges. On the other hand,

$$\int_1^{\infty} \frac{1}{x^2} dx = \lim_{\delta \rightarrow \infty} \left[ -\frac{1}{x} \right]_1^{\delta} = 1,$$

hence the series  $S_2$  converges. □

Notice that the conclusions about convergence remain true even for complex values of  $x$  substituted into the expression. Laurent series can be considered as functions defined on a domain in the complex plain. We return to this in chapter 9. The following theorem is proved already.

CONVERGENCE OF THE LAURENT SERIES ON THE ANNULUS

**Theorem.** *The Laurent series  $S(x)$  centered at  $x_0$  converges for all  $x \in \mathbb{C}$  satisfying  $r < |x - x_0| < R$  and diverges for all  $x$  satisfying  $|x - x_0| < r$  or  $|x - x_0| > R$ , where*

$$r = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_{-n}|}, \quad R^{-1} = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}.$$

The Laurent series need not converge at any point, because possibly  $R < r$ . If we look for an example of the above case of rational functions expanded to Laurent series at some of the roots of the denominator, then clearly  $r = 0$  and therefore, as expected, it converges in the punctured neighbourhood of this point  $x_0$ .  $R$  is given by the distance to the closest root of the denominator. In the case of the first example, the function  $e^{-1/x^2}$ ,  $r = 0$  and  $R = \infty$ .

**6.3.11. Integrals dependent on parameters.**



When integrating a function  $f(x, y_1, \dots, y_n)$  of 1 real variable  $x$  depending on further real parameters  $y_1, \dots, y_n$  with respect to the single variable  $x$ , the result is a function  $F(y_1, \dots, y_n)$  depending on all the parameters. Such a function  $F$  often occurs in practice. For instance, we can look for the volume or area of a body which depends on parameters, and determine the minimal and maximal values (with additional constrains as well). Often it is desirable to interchange the operations of differentiation and integration. That this can be done is proved below. We begin with an examination of continuous dependency on the parameters.

For sake of simplicity, we shall deal with functions  $f(x, y)$  depending on two variables,  $x \in [a, b]$ ,  $y \in [c, d]$ . We say  $f$  is *continuous* on  $I = [a, b] \times [c, d] \subset \mathbb{R}^2 = \mathbb{C}$  if for each  $z = (x, y)$  from the domain of  $f$  and  $\varepsilon > 0$  there is some  $\delta > 0$  such  $|f(w) - f(z)| < \varepsilon$  if  $w \in \mathcal{O}_{\delta}(z)$ . (Notice the definition is the same as with the univariate functions, just we use the distance in the plane.)

The function  $f(x, y)$  is called *uniformly continuous* if for each  $\varepsilon > 0$ , there is  $\delta > 0$  such that for any two points  $z, w$  in  $I \subset \mathbb{R}^2 = \mathbb{C}$ ,  $|z - w| < \delta$  implies  $|f(z) - f(w)| < \varepsilon$ . Exactly the same argument as with univariate functions, based on the fact that every open cover of a compact set in the complex plane contains a finite subcover, cf. Theorem 5.2.8(5), provides the following lemma (cf. the proof of Theorem 6.2.11).

**Lemma.** *Each continuous function  $f(x, y)$  on  $I = [a, b] \times [c, d]$  is uniformly continuous.*

Now we are ready for the following important claim:

**6.C.2.** Using the integral criterion, decide on the convergence of the series

$$\sum_{n=1}^{\infty} \frac{1}{(n+1)\ln^2(n+1)}.$$

**Solution.** The function

$$f(x) = \frac{1}{(x+1)\ln^2(x+1)}, \quad x \in [1, +\infty)$$

is clearly positive and nonincreasing on its whole domain, thus the given series converges if and only if the integral

$$\int_1^{+\infty} f(x) dx$$

converges. By using the substitution  $y = \ln(x+1)$  (where  $dy = dx/(x+1)$ ), we can compute

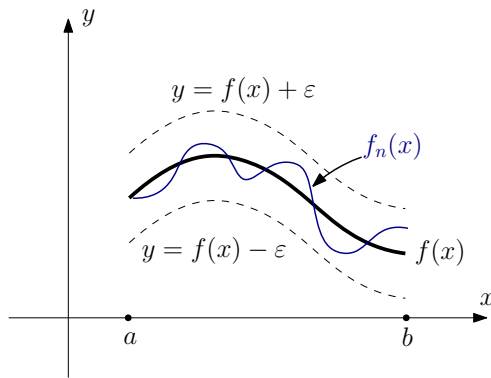
$$\int_1^{+\infty} \frac{1}{(x+1)\ln^2(x+1)} dx = \int_{\ln 2}^{+\infty} \frac{1}{y^2} dy = \frac{1}{\ln 2}.$$

Hence the series converges.  $\square$

Next we will explore examples related to the concept of “uniform convergence”, which is a stronger notion than pointwise convergence.



Recall that given a sequence of functions  $(f_n)$  defined on an interval  $I = [a, b]$ , we say that  $(f_n)$  converges uniformly to a function  $f$  on  $[a, b]$ , if for every  $\varepsilon > 0$  there exists  $N \in \mathbb{N}$  such that for all  $n \geq N$  and  $x \in [a, b]$  we have  $|f_n(x) - f(x)| < \varepsilon$ . In other words, we have  $f(x) - \varepsilon < f_n(x) < f(x) + \varepsilon$  for all  $n \geq N$  and  $x \in [a, b]$ , thus uniform convergence means that all of the functions  $f_n(x)$  are close to  $f(x)$  for all points  $x \in [a, b]$ , except for finitely many of them, see the figure given below and see also 6.3.3.



Notice that we can equivalent interpret the condition of uniform convergence as (see 6.D.46)

$$\lim_{n \rightarrow \infty} \sup_{x \in [a, b]} |f_n(x) - f(x)| = 0.$$

It is also important to note that uniform convergence implies pointwise convergence (and the uniform limit function is the same as the pointwise limit function), but the converse is not true. Before discussing tasks on uniform convergence, we will first emphasize some natural tasks related to the pointwise

**Theorem.** Assume  $f(x, y)$  is a function defined for all  $x$  lying in a bounded interval  $[a, b]$  and all  $y$  in a bounded interval  $[c, d]$ , continuous on  $I = [a, b] \times [c, d]$ . Consider the (Riemann) integral

$$F(y) = \int_a^b f(x, y) dx.$$

Then the function  $F(y)$  is continuous on  $[c, d]$ .

**PROOF.** Fix a point  $y \in [c, d]$ , small  $\varepsilon > 0$ , and choose a neighbourhood  $W$  of  $y$  such that for all  $\bar{y} \in W \subset [c, d]$  and all  $x \in [a, b]$  (remember  $f$  is uniformly continuous)



$$|f(x, \bar{y}) - f(x, y)| < \varepsilon.$$

The Riemann integral of continuous functions is evaluated by approximations of finite sums (equivalently: upper, lower, or Riemann sums with arbitrary representatives  $\xi_i$ , see paragraph 6.2.9).

The goal is to establish that the Riemann sums for the integrals with parameters  $y$  and  $\bar{y}$  cannot differ much. In the following estimate for any partition with  $k$  intervals and representatives  $\xi_i$ , first use the standard properties of the absolute value and then exploit the choice of  $W$ :

$$\begin{aligned} & \left| \sum_{i=0}^{k-1} f(\xi_i, \bar{y})(x_{i+1} - x_i) - \sum_{i=0}^{k-1} f(\xi_i, y)(x_{i+1} - x_i) \right| \\ & \leq \sum_{i=0}^{k-1} |f(\xi_i, \bar{y}) - f(\xi_i, y)|(x_{i+1} - x_i) \\ & < \varepsilon(b - a). \end{aligned}$$

It follows that the limit values for any sequences of the partitions and representatives  $F(y)$  and  $F(\bar{y})$  cannot differ by more than  $\varepsilon(b-a)$  either, so the function  $F$  is continuous.  $\square$

**6.3.12. Integrating twice.** The fact that the integral  $F(y) = \int_a^b f(x, y) dx$  of a continuous function  $f : [a, b] \times [c, d] \rightarrow \mathbb{R}$  in the plane is again a continuous function  $F : [c, d] \rightarrow \mathbb{R}$  allows us to repeat the integration and write



$$(1) \quad I = \int_c^d \int_a^b f(x, y) dx dy = \int_c^d \left( \int_a^b f(x, y) dx \right) dy.$$

The next theorem is the simplest version of the claim known as *Fubini theorem*.

FUBINI THEOREM

**Theorem.** Consider a continuous function  $f : [a, b] \times [c, d] \rightarrow \mathbb{R}$  in the plane  $\mathbb{R}^2$ . The multiple integration (1) is well defined and does not depend on the order of integration, i.e.,

$$I = \int_c^d \left( \int_a^b f(x, y) dx \right) dy = \int_a^b \left( \int_c^d f(x, y) dy \right) dx.$$

convergence of sequences of functions. See also the discussion in 6.3.1, where similar problems are addressed.

**6.C.3.** Provide an example for each of the following scenarios:

- (a) A sequence of continuous functions  $(f_n)$  that converges pointwise to a function  $f$  which is discontinuous;
- (b) A sequence of differentiable functions  $(f_n)$  that converges pointwise to a function  $f$  which is non-differentiable;
- (c) A sequences of integrable functions  $(f_n)$  that converges pointwise to a function  $f$  which is non-integrable.  $\circ$

**6.C.4.** Inspect which of the sequences  $(f_n)_{n \in \mathbb{Z}_+}$  given below, converge uniformly on the given domain:

- (1)  $f_n(x) = \sin(nx)/n$ , with  $x \in \mathbb{R}$ ;
- (2)  $f_n(x) = nx/(n+x)$ , with  $x \in [0, 1]$ ;
- (3)  $f_n(x) = e^{\frac{x^4}{4n^2}}$ , with  $x \in \mathbb{R}$ ;
- (4)  $f_n(x) = e^{\frac{x}{n}}$ , with  $x \in [0, 1]$ ;
- (5)  $f_n(x) = x^n$ , with  $x \in [0, 1]$ .

**Solution.** (1) In the first case it is easy to see that  $f_n \rightarrow 0$  pointwise on  $\mathbb{R}$ , that is, the limit function is the zero one,  $f(x) = 0$ . The sequence  $(f_n)$  also converges uniformly on  $\mathbb{R}$ . This is essentially derived from the inequality

$$|f_n(x)| = \left| \frac{\sin(nx)}{n} \right| = \frac{|\sin(nx)|}{n} \leq \frac{1}{n}.$$

Thus, given some  $\varepsilon > 0$ , we will have  $|f_n(x) - 0| < \varepsilon$  for all  $x \in \mathbb{R}$  if  $n > \frac{1}{\varepsilon}$ . Since  $\frac{1}{\varepsilon}$  does not depend on  $x$  we conclude.

(2) We may express  $f_n$  as

$$f_n(x) = \frac{x}{1 + \frac{x}{n}}, \quad x \in [0, 1],$$

and it is obvious that  $(f_n)$  is pointwise convergent on  $[0, 1]$  with  $\lim_{n \rightarrow \infty} f_n(x) = f(x) = x$ , for all  $x \in [0, 1]$ . Next we see that

$$|f_n(x) - f(x)| = \frac{x^2}{n+x} \leq \frac{1}{n}, \quad x \in [0, 1]$$

and using this we deduce that  $\sup_{x \in [0,1]} |f_n(x) - f(x)| \leq \frac{1}{n}$ . Thus  $\lim_{n \rightarrow \infty} \sup_{x \in [0,1]} |f_n(x) - f(x)| = 0$  and  $(f_n)$  is uniformly convergent in the interval  $[0, 1]$ .

(3) The sequence  $(f_n)_{n \in \mathbb{Z}_+}$  converges pointwise to the constant function  $f(x) = 1$  on  $\mathbb{R}$ , since

$$\lim_{n \rightarrow \infty} e^{\frac{x^4}{4n^2}} = e^0 = 1, \quad x \in \mathbb{R}.$$

However, we see that  $f_n(\sqrt{2n}) = e > 2$ , for all  $n \in \mathbb{Z}_+$ , which does not allow uniform convergence over  $\mathbb{R}$  (be aware that in the definition of uniform convergence, it suffices to consider  $\varepsilon \in (0, 1)$ ).

(4) Obviously,  $\lim_{n \rightarrow \infty} e^{\frac{x}{n}} = 1$ , i.e., the pointwise limit function is the constant function  $f(x) = 1$ . Next we see that

$$\sup_{x \in [0,1]} |f_n(x) - f(x)| = \sup_{x \in [0,1]} (e^{\frac{x}{n}} - 1) = e^{\frac{1}{n}} - 1,$$

**PROOF.** We know  $f$  is uniformly continuous on the product of intervals  $[a, b] \times [c, d]$  in the plane. Thus, for each  $\varepsilon > 0$  there is  $\delta > 0$  such that  $|f(x_1, y_1) - f(x_2, y_2)| < \varepsilon$  whenever  $|x_1 - x_2| < \delta$  and  $|y_1 - y_2| < \delta$ .



We know both Riemann integrals in (1) exist, thus we may fix a sequence  $\Xi_k$  of partitions of the interval  $[a, b]$  into  $k$  subinterval  $[x_{i-1}, x_i]$  of equal size  $1/k$  and with representatives  $\xi_{i,k}$ ,  $i = 1, \dots, k$ , and similarly for the interval  $[c, d]$  with the subintervals  $[y_{j-1}, y_j]$  and representatives  $\eta_{j,k}$ ,  $j = 1, \dots, k$ . Then we may write

$$I = \int_c^d \lim_{k \rightarrow \infty} S_k(y) dy, \quad S_k(y) = \left( \sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b-a) \right)$$

If  $\frac{1}{k} < \delta$ , then

$$\begin{aligned} & \left| \int_a^b f(x, y) dx - \sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b-a) \right| \\ & \leq \sum_{i=1}^k \left| \int_{x_{i-1}}^{x_i} f(x, y) dx - f(\xi_{i,k}, y) \frac{1}{k} (b-a) \right| \\ & \leq \sum_{i=1}^k \int_{x_{i-1}}^{x_i} |f(x, y) - f(\xi_{i,k}, y)| dx \\ & \leq \varepsilon (b-a). \end{aligned}$$

Thus, the convergence of

$$\lim_{k \rightarrow \infty} S_k(y) = F(y) = \int_a^b f(x, y) dx$$

is uniform on  $[c, d]$ . In particular, we may swap the integral and the limit to obtain

$$\begin{aligned} I &= \lim_{k \rightarrow \infty} \int_c^d \left( \sum_{i=1}^k f(\xi_{i,k}, y) \frac{1}{k} (b-a) \right) dy \\ &= \lim_{k \rightarrow \infty} \sum_{i=1}^k \left( \int_c^d f(\xi_{i,k}, y) dy \right) \frac{1}{k} (b-a) \\ &= \lim_{k \rightarrow \infty} \left( \lim_{\ell \rightarrow \infty} \sum_{i=1}^k \sum_{j=1}^{\ell} f(\xi_{i,k}, \eta_{j,\ell}) \frac{1}{k} \frac{1}{\ell} (b-a)(d-c) \right). \end{aligned}$$

This double limit can be clearly rewritten as

$$\lim_{k \rightarrow \infty} \left( \sum_{i=1}^k \sum_{j=1}^{\ell_k} f(\xi_{i,k}, \eta_{j,\ell_k}) \frac{1}{k} \frac{1}{\ell_k} (b-a)(d-c) \right)$$

for a suitable sequence of indices  $\ell_k \rightarrow \infty$ .

Finally, take any sequence of indices  $k_n \rightarrow \infty$  and  $\ell_n \rightarrow \infty$ , divide intervals  $(a, b)$  and  $(c, d)$  to  $k_n$  and  $\ell_n$  equal parts, and choose some representatives  $(x_{i,k_n}, y_{j,\ell_n})$  in all the corresponding small rectangles. Then the absolute value of the difference

$$\sum_{i=1}^{k_n} \sum_{j=1}^{\ell_n} f(\xi_{i,k_n}, \eta_{j,\ell_n}) - \sum_{i=1}^{k_n} \sum_{j=1}^{\ell_n} f(x_{i,k_n}, y_{j,\ell_n})$$

and hence

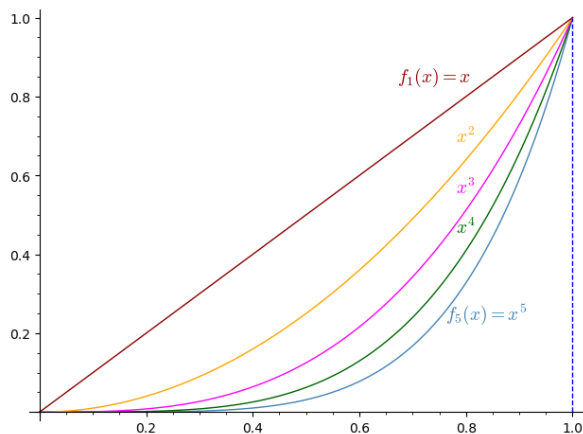
$$\lim_{n \rightarrow \infty} \sup_{x \in [0,1]} |f_n(x) - f(x)| = \lim_{n \rightarrow \infty} (e^{\frac{1}{n}} - 1) = 1 - 1 = 0.$$

Thus  $(f_n)$  is uniformly convergent over  $[0, 1]$ .

(5) Let us first use Sage to sketch some members of  $(f_n)$ . We use the following cell:

```
q=plot(x^5, x, 0, 1, color="steelblue")
q+= text (r"$f_5(x)=x^5$", (0.84, 0.25),
color="steelblue", fontsize="14")
#q+=plot(x^5, x, 0, pi, color="black")
q+=plot(x^4, x, 0, 1, color="darkgreen")
q+= text (r"$x^4$", (0.8, 0.47),
color="darkgreen", fontsize="14")
q+=plot(x^3, x, 0, 1, color="magenta")
q+= text (r"$x^3$", (0.8, 0.57),
color="magenta", fontsize="14")
q+=plot(x^2, x, 0, 1, color="orange")
q+= text (r"$x^2$", (0.8, 0.7),
color="orange", fontsize="14")
q+=plot(x, x, 0, 1, color="darkred")
q+= text (r"$f_1(x)=x$", (0.74, 0.85),
color="darkred", fontsize="14")
q+=line([(1, 0), (1, 1^5)], linestyle="--")
q.show(ymax=1)
```

Executing this block we obtain the graph of the first 5 members of the sequence  $(f_n)$ , which we present here:



An alternative but less informative way to plot some terms of  $(f_n)$  goes as follows:

```
var("n"); f(n, x)=x^n
p=plot([f(n, x) for n in [1, 2..5]], (x, 0, 1))
show(p)
```

As we have already seen in Chapter 5 (see 5.B.3), the limit function of  $(f_n(x) = x^n)$  on  $[0, 1]$  is given by

$$\lim_{n \rightarrow \infty} (x^n) = f(x) = \begin{cases} 0, & \text{if } x \in [0, 1), \\ 1, & \text{if } x = 1. \end{cases}$$

Thus

$$|f_n(x) - f(x)| = \begin{cases} 0, & \text{for } x = 1, \\ |x^n|, & \text{for } x \in [0, 1), \end{cases}$$

is at most  $\varepsilon(b-a)(c-d)$  whenever  $\frac{1}{k_n} < \delta$  and  $\frac{1}{\ell_n} < \delta$ . Thus, we see that the value of the iterated integral can be expressed as the limit of double sums, where the representatives of the individual rectangles are chosen arbitrarily.<sup>8</sup> This expression does not depend on the order of our integration and thus the two possibilities for the iterated integrals must coincide.  $\square$

**6.3.13. Differentiation in the integrals.** We are ready to discuss the differentiation of integrals with respect to parameters. The following result is extremely useful. For instance we shall use it in the next chapter when examining integral transforms.



DIFFERENTIATION WITH RESPECT TO PARAMETERS

**Theorem.** Consider a continuous function  $f(x, y)$  defined for all  $x$  from a finite interval  $[a, b]$  and for all  $y$  in another finite interval  $[c, d]$ , a point  $z \in [c, d]$ , and the integral

$$F(y) = \int_a^b f(x, y) dx.$$

If there exists the continuous derivative  $\frac{d}{dy} f$  on a neighbourhood of the point  $z$ , then  $\frac{d}{dy} F(z)$  exists as well and

$$\frac{d}{dy} F(z) = \int_a^b \frac{d}{dy} f(x, y)|_{y=z} dx.$$

**PROOF.** By the assumed continuity of all functions and the already known continuous dependence of integrals on parameters, some knowledge about univariate antiderivatives can be used. The result is then a simple consequence of the Fubini theorem. Denote

$$G(y) = \int_a^b \frac{d}{dy} f(x, y) dx, \quad F(y) = \int_a^b f(x, y) dx$$

and compute, invoking Fubini theorem, the antiderivative

$$\begin{aligned} H(y) &= \int_{y_0}^y G(z) dz = \int_{y_0}^y \left( \int_a^b \frac{d}{dz} f(x, z) dx \right) dz \\ &= \int_a^b \left( \int_{y_0}^y \frac{d}{dz} f(x, z) dz \right) dx \\ &= \int_a^b (f(x, y) - f(x, y_0)) dx \\ &= F(y) - F(y_0). \end{aligned}$$

Finally, differentiating with respect to  $y$  yields

$$G(y) = \frac{d}{dy} H(y) = \frac{dF}{dy}(y),$$

as desired.  $\square$

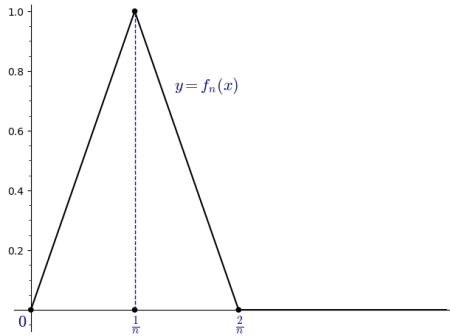
<sup>8</sup>Actually, this is the way how we shall define the Riemann integral in more variables in Chapter 8.

which implies that

$$\sup_{x \in [0,1]} |f_n(x) - f(x)| = \sup\{1, 0\} = 1.$$

Therefore,  $\lim_{n \rightarrow \infty} \sup_{x \in [0,1]} |f_n(x) - f(x)| = 1 \neq 0$ , and  $(x^n)_{n \in \mathbb{Z}_+}$  is not uniformly convergent on  $[0, 1]$ .  $\square$

**6.C.5.** Demonstrate that the sequence  $(f_n)_{n \in \mathbb{Z}_+}$  of non-negative functions defined by the given graph, pointwise converges to 0, yet it does not converge uniformly on  $[0, +\infty)$ .



**6.C.6.** Show that the sequence  $(f_n)_{n \in \mathbb{Z}_+}$  of functions defined by

$$f_n : \mathbb{R} \rightarrow \mathbb{R}, \quad f_n(x) = \sqrt{x^2 + \frac{1}{n^2}},$$

converges uniformly on  $\mathbb{R}$ .

**6.C.7.** (a) Provide an example of sequences  $(f_n), (g_n)$  that converge uniformly on a set  $I$ , but the product sequence  $(f_n \cdot g_n)$  does not converge uniformly on  $I$ .

(b) Suppose that the sequences  $(f_n), (g_n)$  converge uniformly to  $f, g$ , respectively, on a set  $I$ . Assume also that there exists some  $M > 0$  such that  $|f(x)| < M$  and  $|g(x)| < M$  for all  $x \in I$ . Show that the sequence  $(f_n \cdot g_n)$  uniformly converges to  $f \cdot g$  on  $I$ .

**6.C.8.** Inspect which of the sequences  $(f_n)_{n \in \mathbb{Z}_+}$  given below, converge uniformly on the given domain.

*Similarly with uniform convergence of sequences of functions, we can explore series of functions which converge uniformly on an interval. This is the case when the sequence of the partial sums converges uniformly. Uniform convergence of series has many applications. For instance, in Chapter 7 we will discuss the uniform convergence of trigonometric series, specifically focusing on Fourier series.*

**6.C.9.** Decide whether the series

$$\sum_{n=1}^{\infty} \frac{\sqrt{x} \cdot n}{n^4 + x^2}$$

converges uniformly on the interval  $(0, +\infty)$ .

**Solution.** Using the denotation

**6.3.14. The Riemann–Stieltjes integral.** To end this chapter, we mention briefly some other concepts of integration. Mostly we confine ourselves to remarks and comments. Readers interested in a thorough explanation can find another source.



First, a modification of the Riemann integral, which is useful when discussing probability and statistics. In the discussion of integration, we summed infinitely many linearized (infinitely) small increments of the area given by a function  $f(x)$ . We omitted the possibility that for different values of  $x$  we could take the increments with different weights. This can be arranged at the infinitesimal level by exchanging the differential  $dx$  for  $\varphi(x)dx$  for some suitable function  $\varphi$ , or we could even take some single points as adding the weight independently of the size of our intervals in the partitions.

Imagine that at some point  $x_0$ , the increment of the integrated quantity is given by  $\alpha f(x_0)$  independently of the size of the increment of  $x$ . For example, we may observe the probability that the amount of alcohol per mille in the blood of a driver at a test will be at most  $x$ . We might like to integrate over the possible values in the interval  $[0, x]$ . With quite a large probability the value is 0. Thus for any integral sum, the segment containing zero contributes by a constant nonzero contribution, independent of the norm of the partition.

We cannot simulate such behaviour by multiplying the differential  $dx$  by some real function. Instead we generalize the Riemann integral in the following way:

#### RIEMANN–STIELTJES INTEGRAL

Choose a real (usually) nondecreasing function  $g$  on a finite interval  $[a, b]$ . For every partition  $\Xi$  with representative  $\xi_i$  and points of the partition

$$a = x_0, x_1, \dots, x_n = b$$

the Riemann–Stieltjes integral sum of function  $f(x)$  is

$$S_{\Xi} = \sum_{i=1}^n f(\xi_i)(g(x_i) - g(x_{i-1})).$$

The Riemann–Stieltjes integral

$$I = \int_a^b f(x)dg(x)$$

exists and its value is  $I$ , if for every real  $\varepsilon > 0$  there exists a norm of the partition  $\delta > 0$  such that for all partitions  $\Xi$  with norm smaller than  $\delta$ ,

$$|S_{\Xi} - I| < \varepsilon.$$

For example, choose  $g(x)$  on interval  $[0, 1]$  as a piecewise constant function with finitely many discontinuities  $c_1, \dots, c_k$  and “jumps”

$$\alpha_i = \lim_{x \rightarrow c_i+} g(x) - \lim_{x \rightarrow c_i-} g(x),$$

$$f_n(x) = \frac{\sqrt{x} \cdot n}{n^4 + x^2}, \quad x > 0, \quad n \in \mathbb{N},$$

we have

$$f'_n(x) = \frac{n(n^4 - 3x^2)}{2\sqrt{x}(n^4 + x^2)^2}, \quad x > 0, \quad n \in \mathbb{N}.$$

From now on, let  $n \in \mathbb{N}$  be arbitrary. The inequalities  $f'_n(x) > 0$  for  $x \in (0, n^2/\sqrt{3})$  and  $f'_n(x) < 0$  for  $x \in (n^2/\sqrt{3}, +\infty)$  imply that the maximum of function  $f_n$  is attained exactly at the point  $x = n^2/\sqrt{3}$ . Since

$$f_n\left(\frac{n^2}{\sqrt{3}}\right) = \frac{\sqrt[4]{27}}{4n^2} \quad \text{a} \quad \sum_{n=1}^{\infty} \frac{\sqrt[4]{27}}{4n^2} = \frac{\sqrt[4]{27}}{4} \sum_{n=1}^{\infty} \frac{1}{n^2} < +\infty,$$

according to the Weierstrass test, the series  $\sum_{n=1}^{\infty} f_n(x)$  converges uniformly on the interval  $(0, +\infty)$ .  $\square$

**6.C.10.** For  $x \in [-1, 1]$ , add

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n(n+1)} x^{n+1}.$$

**Solution.** First notice that by the symbol for an indefinite integral, we'll denote one specific primitive function (while preserving the variable), which should be understood as a so called function of the upper limit, while the lower limit is zero. Using the theorem about integration of a power series for  $x \in (-1, 1)$ , we'll obtain

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n(n+1)} x^{n+1} &= \sum_{n=1}^{\infty} \left( \frac{(-1)^{n+1}}{n} \int x^n dx \right) \\ &= \int \sum_{n=1}^{\infty} \left( \frac{(-1)^{n+1}}{n} x^n \right) dx \\ &= \int \sum_{n=1}^{\infty} ((-1)^{n+1} \int x^{n-1} dx) dx \\ &= \int \left( \int \sum_{n=1}^{\infty} (-x)^{n-1} dx \right) dx = \int \left( \int 1 - x + x^2 - x^3 + \dots dx \right) dx \\ &= \int \left( \int \frac{1}{1+x} dx \right) dx = \int \ln(1+x) + C_1 dx. \end{aligned}$$

Since

$$\int \sum_{n=1}^{\infty} \left( \frac{(-1)^{n+1}}{n} x^n \right) dx = \int \ln(1+x) + C_1 dx,$$

we know from the continuity of the given functions that

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = \ln(1+x) + C_1, \quad x \in (-1, 1).$$

The choice  $x = 0$  then yields  $0 = \ln 1 + C_1$ , i.e.  $C_1 = 0$ . Next,

$$\begin{aligned} \int \ln(1+x) dx &= \text{per partes} \\ &= \left| \begin{array}{ll} u = \ln(1+x) & u' = \frac{1}{1+x} \\ v' = 1 & v = x \end{array} \right| \\ &= x \ln(1+x) - \int \frac{x}{1+x} dx = x \ln(1+x) - \int 1 - \frac{1}{1+x} dx \\ &= x \ln(1+x) - x + \ln(1+x) + C_2 \\ &= (x+1) \ln(x+1) - x + C_2. \end{aligned}$$

Since the given series converges at the point  $x = 0$  with a sum of 0, analogously as for  $C_1$ ,

$$0 = 1 \cdot \ln 1 - 0 + C_2$$

implies that  $C_2 = 0$ . In total, we have for  $x \in (-1, 1)$ :

then the Riemann–Stieltjes integral exists for every continuous  $f(x)$  and equals

$$I = \int_0^1 f(x) dg(x) = \sum_{i=1}^k \alpha_i f(c_i).$$

By the same technique as used for the Riemann integral, we define upper and lower sums and upper and lower Riemann–Stieltjes integral. For bounded functions they always exist, and their values coincide if and only if the Riemann–Stieltjes integral in the above sense exists.

We have already encountered problems with the Riemann integration of functions that are “too jumpy”. For a function  $g(x)$  on a finite interval  $[a, b]$  define its *variation* by

$$\text{var}_a^b g = \sup_{\Xi} \sum_{i=1}^n |g(x_i) - g(x_{i-1})|,$$

where the supremum is taken over all partitions  $\Xi$  of the interval  $[a, b]$ . If the supremum is infinite, we say that  $g(x)$  has unbounded variation on  $[a, b]$ . Otherwise we say that  $g$  is a function with bounded variation on  $[a, b]$ . A function is of bounded variation if and only if it can be written as the difference of two monotonic functions. We shall not provide a proof here.

As in the discussion of the Riemann integral, we derive the following theorem.

PROPERTIES OF THE RIEMANN–STIELTJES INTEGRAL

**Theorem.** Let  $f(x)$  and  $g(x)$  be real functions on a finite interval  $[a, b]$ .

- (1) Suppose  $g(x)$  is non-decreasing and continuously differentiable. Then the Riemann integral on the left hand side and the Riemann–Stieltjes integral on the right hand side either both exist or do not exist. In the former case, their values are equal

$$\int_a^b f(x)g'(x)dx = \int_a^b f(x)dg(x)$$

- (2) If  $f(x)$  is continuous and  $g(x)$  is a function with finite variation, then the integral  $\int_a^b f(x)dg(x)$  exists.

We invite the reader to add the details of its proof. The main tools are the mean theorem, the uniform continuity of continuous functions on closed bounded intervals. The variation of  $g$  over the interval  $[a, b]$  plays the role of the length of the interval in the earlier proofs dealing with Riemann integration.



**6.3.15. Kurzweil–Henstock integral.** The last topic in this chapter is a modification of the Riemann integral, which fixes the unfortunate behaviour at the third point in the paragraph 6.3.1. That is, the limits of the non-decreasing sequences of integrable functions are again integrable. Then we can interchange the order





$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n(n+1)} x^{n+1} = (x+1) \ln(x+1) - x.$$

Moreover, according to Abel's theorem (see 6.3.9), the sum of the given series equals the (potentially improper) limit of the function  $(x+1) \ln(x+1) - x$  at points  $-1$  and  $1$ . In our case, both limits are proper (at point  $1$ , the function is even continuous and the value of the limit at point  $1$  then equals the value of the function  $2 \ln 2 - 1$ ). For computing the value of the limit at point  $-1$ , we'll use L'Hospital's rule:

$$\begin{aligned} \lim_{x \rightarrow -1^+} (x+1) \ln(x+1) - x &= \lim_{t \rightarrow 0^+} t \ln t + 1 \\ &= \lim_{t \rightarrow 0^+} \frac{\ln t}{\frac{1}{t}} + 1 = \lim_{t \rightarrow 0^+} \frac{\frac{1}{t}}{-\frac{1}{t^2}} + 1 \\ &= \lim_{t \rightarrow 0^+} -t + 1 = 1. \end{aligned}$$

Of course, the convergence of the series at points  $\pm 1$  can be verified directly. It's even possible to directly deduce that  $\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$  (by writing out  $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$ ).  $\square$

**6.C.11. Sum of a series.** Using theorem 6.3.5 "about the interchange of a limit and an integral of a sequence of uniformly convergent functions", we'll now add the number series

$$\sum_{n=1}^{\infty} \frac{1}{n2^n}.$$

We'll use the fact that  $\int_2^{\infty} \frac{dx}{x^{n+1}} = \frac{1}{n2^n}$ .

**Solution.** On interval  $(2, \infty)$ , the series of functions  $\sum_{n=1}^{\infty} \frac{1}{x^{n+1}}$  converges uniformly. That is implied for example by the Weierstrass test: each of the function  $\frac{1}{x^{n+1}}$  is decreasing on interval  $(2, \infty)$ , thus their values are at most  $\frac{1}{2^{n+1}}$ ; the series  $\sum_{n=1}^{\infty} \frac{1}{2^{n+1}}$  is convergent though (it's a geometric series with quotient  $\frac{1}{2}$ ). Hence according to the Weierstrass test, the series of functions  $\sum_{n=1}^{\infty} \frac{1}{x^{n+1}}$  converges uniformly. We can even write the resulting function explicitly. Its value at any  $x \in (2, \infty)$  is the value of the geometric series with quotient  $\frac{1}{x}$ , so if we denote the limit by  $f(x)$ , we have

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{x^{n+1}} = \frac{1}{x^2} \frac{1}{1 - \frac{1}{x}} = \frac{1}{x(x-1)}.$$

By using (6.3.7) (3), we get

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n2^n} &= \sum_{n=1}^{\infty} \int_2^{\infty} \frac{dx}{x^{n+1}} = \int_2^{\infty} \left( \sum_{n=1}^{\infty} \frac{1}{x^{n+1}} \right) dx \\ &= \int_2^{\infty} \frac{1}{x(x-1)} dx = \lim_{\delta \rightarrow \infty} \int_2^{\delta} \frac{1}{x-1} - \frac{1}{x} dx \\ &= \lim_{\delta \rightarrow \infty} [(\ln(\delta-1) - \ln(\delta) - \ln(1) + \ln 2)] \\ &= \lim_{\delta \rightarrow \infty} \left[ \ln \left( \frac{\delta-1}{\delta} \right) \right] + \ln(2) \\ &= \ln \left( \lim_{\delta \rightarrow \infty} \frac{\delta-1}{\delta} \right) + \ln 2 = \ln 2 \end{aligned}$$

of the limit process and integration in these cases, just as with uniform convergence.

Notice what is the essence of the problem. Intuitively we assume that very small sets must have zero size. Thus the changes of values of the functions on such sets should not change the integral. Moreover, a countable union of such sets which are "negligible for the purpose of integration" should also have zero size. We would expect for example that the set of rational numbers inside a finite interval would have this property, hence its characteristic function should be integrable and the value of such an integral should be zero.

We say that a set  $A \subset \mathbb{R}$  has *zero measure*, if for every  $\varepsilon > 0$  there is a covering of the set  $A$  by a countable system of open intervals  $J_i, i = 1, 2, \dots$  such that

$$\sum_{i=1}^{\infty} m(J_i) < \varepsilon.$$

$m(J_i)$  means the length of the interval  $J_i$ .

In the sequel, the statement "function  $f$  has the given property on a set  $B$  almost everywhere" means that  $f$  has this property at all points except for a subset  $A \subset B$  of zero measure. For example, the characteristic function of rational numbers is zero almost everywhere. A piece-wise continuous function is continuous almost everywhere.

Now we modify the definition of the Riemann integral so that restrictions on the Riemann sums are permitted, eliminating the effect of the values of the integrated function on sets of measure zero. This is achieved by a finer control of the size of the segments in the partition in the vicinity of problematic points.

A positive real function  $\delta$  on a finite interval  $[a, b]$  is called a *gauge*. A partition  $\Xi$  of interval  $[a, b]$  with representatives  $\xi_i$  is  $\delta$ -gauged, if

$$\xi_i - \delta(\xi_i) < x_{i-1} \leq \xi_i \leq x_i < \xi_i + \delta(\xi_i)$$

for all  $i$ .

The norm  $\delta$  of the partition used in the Riemann integration is a special case of constant gauges  $\delta(x) = \delta > 0$ . In order to restrict the Riemann sums to a gauged partition with representatives in the definition of the integral, it is necessary to know that for every gauge  $\delta$ , a  $\delta$ -gauged partition with representatives exists. Otherwise the condition in the definition could be satisfied in a vacuous way. This statement is called *Cousin's lemma*. It is proved by exploiting the standard properties of suprema:

For a given gauge  $\delta$  on  $[a, b]$ , denote by  $M$  the set of all points  $x \in [a, b]$  such that a  $\delta$ -gauged partition with representatives can be found on  $[a, x]$ .  $M$  is nonempty and bounded, thus it has a supremum  $s$ . If  $s \neq b$ , then there is a gauged partition with representatives at  $s$ , where  $s$  is in the interior of the last segment. This leads to a contradiction. Thus the supremum is  $b$ , but then the gauge  $\delta(b) > 0$  and thus  $b$  itself belongs to the set  $M$ .

**6.C.12.** Consider function  $f(x) = \sum_{n=1}^{\infty} ne^{-nx}$ . Determine

$$\int_{\ln 2}^{\ln 3} f(x) dx.$$

**Solution.** Similarly as in the previous case, the Weierstrass test for uniform convergence implies that the series of functions  $\sum_{n=1}^{\infty} ne^{-nx}$  converges uniformly on interval  $(\ln 2, \ln 3)$ , since each of the functions  $ne^{-nx}$  is lesser than  $\frac{n}{2^n}$  on  $(\ln 2, \ln 3)$  and the series  $\sum_{n=1}^{\infty} \frac{n}{2^n}$  converges, which can be seen for example from the ratio test for convergence of series:

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n \rightarrow \infty} \frac{(n+1)2^{-(n+1)}}{n2^n} = \lim_{n \rightarrow \infty} \frac{1}{2} \frac{n+1}{n} = \frac{1}{2}.$$

In total, according to (6.3.7) (3), we have

$$\begin{aligned} \int_{\ln 2}^{\ln 3} f(x) dx &= \int_{\ln 2}^{\ln 3} \sum_{n=1}^{\infty} ne^{-nx} dx = \sum_{n=1}^{\infty} \int_{\ln 2}^{\ln 3} ne^{-nx} dx \\ &= \sum_{n=1}^{\infty} [-e^{-nx}]_{\ln 2}^{\ln 3} = \sum_{n=1}^{\infty} \left( \frac{1}{2^n} - \frac{1}{3^n} \right) = 1 - \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

**6.C.13.** Determine the following limit (give reasons for the procedure of computation):

$$\lim_{n \rightarrow \infty} \int_0^{\infty} \frac{\cos\left(\frac{x}{n}\right)}{\left(1 + \frac{x}{n}\right)^n} dx.$$

**Solution.** First we'll determine  $\lim_{n \rightarrow \infty} \frac{\cos\left(\frac{x}{n}\right)}{\left(1 + \frac{x}{n}\right)^n}$ . The sequence of these functions converges pointwise and we have

$$\lim_{n \rightarrow \infty} \frac{\cos\left(\frac{x}{n}\right)}{\left(1 + \frac{x}{n}\right)^n} = \frac{1}{\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n} \stackrel{??}{=} \frac{1}{e^x}$$

It can be shown that the given sequence converges uniformly. Then according to (6.3.5),

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_0^{\infty} \frac{\cos\left(\frac{x}{n}\right)}{\left(1 + \frac{x}{n}\right)^n} dx &= \int_0^{\infty} \left[ \lim_{n \rightarrow \infty} \frac{\cos\left(\frac{x}{n}\right)}{\left(1 + \frac{x}{n}\right)^n} \right] dx \\ &= \int_0^{\infty} \frac{1}{e^x} dx = 1 \end{aligned}$$

We leave the verification of uniform convergence to the reader (we only point out that the discussion is more complicated than in the previous cases).

**6.C.14.** Find the analytic function whose Taylor series is

$$x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \dots,$$

for  $x \in [-1, 1]$ .

**6.C.15.** From the knowledge of the sum of a geometric series, derive the Taylor series of function

$$y = \frac{1}{5+2x}$$

Now we can state the following generalization of the Riemann integral. Call it the *K-integral*<sup>9</sup>.

KURZWEIL-HENSTOCK INTEGRAL

**Definition.** A function  $f$  defined on a finite interval  $[a, b]$  has a *Kurzweil-Henstock integral*

$$I = \int_a^b f(x) dx,$$

if for every  $\varepsilon > 0$ , there exists a gauge  $\delta$  such that for every  $\delta$ -gauged partition with representatives  $\Xi$ , the inequality  $|S_{\Xi} - I| < \varepsilon$  is true for the corresponding Riemann sum  $S_{\Xi}$ .

**6.3.16. Basic properties.** When defining the K-integral, only the set of all partitions is bounded, for which the Riemann sums are taken into account. Hence if the function is Riemann integrable, then it is K-integrable, and the two integrals are equal.

For the same reason, the argumentation in Theorem 6.2.8 about simple properties of the Riemann integral applies. This verifies that the K-integral behaves in the same way. In particular, a linear combination of integrable function  $cf(x) + dg(x)$  is again integrable and its integral is  $c \int_a^b f(x) dx + d \int_a^b g(x) dx$  etc. To prove this, it suffices only to think through some modifications when discussing the refined partitions, which moreover should be  $\delta$ -gauged.

The Kurzweil integral behaves as anticipated with respect to the sets of zero measure:

**Theorem.** Consider a function  $f$ , defined on the interval  $[a, b]$ . If  $f$  is zero almost every, then the K-integral  $\int_a^b f(x) dx$  exists and is zero.

**PROOF.** The proof is an illustration of the idea that the influence of values on a "small" set can be removed by a suitable choice of gauge. Denote by  $M$  the corresponding set of zero measure, outside of which  $f(x) = 0$  and write  $M_k \subset [a, b]$ ,  $k = 1, \dots$ , for the subset of the points for which  $k - 1 < |f(x)| \leq k$ . Because all the sets  $M_k$  have zero measure, each of them can be covered by a countable system of pairwise disjoint open intervals  $J_{k,i}$  such that the sum of their lengths is arbitrarily small.

Define the gauge  $\delta(x)$  for  $x \in J_{k,i}$  so that the intervals  $(x - \delta(x), x + \delta(x))$  are still contained in  $J_{k,i}$ . Outside of  $M$ ,  $\delta$  is defined arbitrarily.

<sup>9</sup>There are many equivalent definitions and thus also names for this K-integral. A complicated approach was coined by Arnaud Denjoy around 1912. Thus the space of real functions integrable on an interval  $[a, b]$  in this sense is often called *Denjoy space*. Other people involved were Nikolai Luzin and Oskar Perron. We can find the integral under their names. The simple and beautiful definition was introduced by Jaroslav Kurzweil, a Czech mathematician still living in 1957. Much of the theory was developed by Ralph Henstock (1923-2007), an English mathematician.

centered at the origin. Then determine its radius of convergence. ○

**6.C.16.** Expand the function

$$y = \frac{1}{3-2x}, \quad x \in \left(-\frac{3}{2}, \frac{3}{2}\right)$$

to a Taylor series centered at the origin. ○

**6.C.17.** Expand the function  $\cos^2(x)$  to a power series at the point  $\pi/4$  and determine for which  $x \in \mathbb{R}$  this series converges. ○

**6.C.18.** Express the function  $y = e^x$  defined on the whole real axis as an infinite polynomial with terms of the form  $a_n(x-1)^n$  and express the function  $y = 2^x$  defined on  $\mathbb{R}$  as an infinite polynomial with terms  $a_n x^n$ . ○

**6.C.19.** Find a function  $f$  such that for  $x \in \mathbb{R}$ , the sequence of functions

$$f_n(x) = \frac{n^2 x^3}{n^2 x^2 + 1}, \quad n \in \mathbb{N}$$

to it. Is this convergence uniform on  $\mathbb{R}$ ? ○

**6.C.20.** Does the series

$$\sum_{n=1}^{\infty} \frac{n x}{n^4 + x^2}, \quad \text{kde } x \in \mathbb{R},$$

converge uniformly on the whole real axis? ○

**6.C.21.** By using differentiation, obtain the Taylor expansion of function  $y = \cos x$  from the Taylor expansion of function  $y = \sin x$  centered at the origin. ○

**6.C.22.** Approximate

- (a) cosine of ten degrees with a precision of at least  $10^{-5}$ ;
  - (b) the definite integral  $\int_0^{1/2} \frac{dx}{x^4+1}$  with a precision of at least  $10^{-3}$ .
- 

**6.C.23.** Determine the power expansion centered at  $x_0 = 0$  of function

$$f(x) = \int_0^x e^{t^2} dt, \quad x \in \mathbb{R}.$$
○

**6.C.24.** Find the analytic function whose Taylor series is

$$x - \frac{1}{3} x^3 + \frac{1}{5} x^5 - \frac{1}{7} x^7 + \dots,$$

for  $x \in [-1, 1]$ . ○

**6.C.25.** From the knowledge of the sum of a geometric series, derive the Taylor series of function

$$y = \frac{1}{5+2x}$$

centered at the origin. Then determine its radius of convergence. ○

For any  $\delta$ -gauged partition  $\Xi$  of the interval  $[a, b]$  the bound on the corresponding Riemann sum is given as

$$\begin{aligned} \left| \sum_{j=0}^{n-1} f(\xi_n)(x_{i+1} - x_i) \right| &= \left| \sum_{\substack{j=0 \\ \xi_i \in M}}^{n-1} f(\xi_n)(x_{i+1} - x_i) \right| \\ &\leq \sum_{k=1}^{\infty} \sum_{\substack{j=0 \\ \xi_i \in M_k}}^{n-1} |f(\xi_n)|(x_{i+1} - x_i) \\ &\leq \sum_{k=1}^{\infty} k \left( \sum_{\substack{j=0 \\ \xi_i \in M_k}}^{n-1} m(J_{k,j}) \right). \end{aligned}$$

To guarantee that this bound is smaller than a given  $\varepsilon$ , it suffices to choose the covering by the intervals  $J_{k,j}$  so that

$$\sum_{j=1}^{\infty} m(J_{k,j}) \leq \frac{\varepsilon}{k2^k}.$$

Since  $\sum_{k=1}^{\infty} 2^{-k} = 1$ , the result follows. □

**Corollary.** *If the values of  $f(x)$  are changed on a set of zero measure, the  $K$ -integrability of  $f(x)$  is not changed, and neither is the value of its integral.*

**6.3.17. The fundamental theorems of Calculus.** We conclude this chapter with a few remarks on the properties of integration procedures from the point of view of expectations and reality.<sup>10</sup>

In 6.2.9, we deal with the relation between the derivatives  $f(t)$  and the antiderivatives (integrals)  $F(t)$ . Since  $f(t)$  is assumed continuous, two essential claims collapse into one, resulting in

$$F(t) = \int_{t_0}^t f(x) dx$$

up to the choice of the value of  $F(t_0)$ . In particular,

$$\int_{t_0}^t F'(dx) dx = F(t) - f(t_0)$$

for all choices of  $F$ .

More generally, this can be split into two claims which hold for the  $K$ -integral under much milder conditions:

<sup>10</sup>A very good and elementary exposition of the  $K$ -integral can be found in the short paper *Return to the Riemann Integral*. The American Mathematical Monthly, Vol. 103, No. 8 (1996), 625-632. by Robert G. Bartle.

**Theorem.** (1) Suppose the K-integral  $\int_a^b f(x) dx$  exists. Then  $F(t) = \int_a^t f(x) dx$  is continuous. The derivative  $F'(t)$  exists and equals  $f(t)$  almost everywhere.  
 (2) Suppose  $F(t)$  is a continuous function on  $[a, b]$  and suppose  $f(t) = F'(t)$  exists for all but countably many exceptional points  $t$  in  $[a, b]$ . If  $f(t)$  is defined arbitrarily at those points, then  $F(t) = \int_a^t f(x) dx$  exists for all  $t \in [a, b]$  and equals to  $F(t) - F(a)$ .

We have no space here to go into proofs. Notice, the statements of the theorem are characteristic for the K-integrals, i.e. this is the only integration concept on  $\mathbb{R}$  for which these two theorems hold true.

**6.3.18. K-integrability and Lebesgue measure.** We illustrate the claims in the latter theorem on the indicator function  $\chi_{\mathbb{Q}}$  of the rational numbers. Clearly its K-integral

$$F(t) = \int_a^t \chi_{\mathbb{Q}}(x) dx$$

exists ( $\chi_{\mathbb{Q}}$  is zero almost everywhere) and equals zero. Its derivative  $F'(t)$  is identically zero, and equals  $\chi_{\mathbb{Q}}$  nearly everywhere. This is a good example of a bounded function which is not Riemann integrable, but is integrable in the more general sense.

There are many more K-integrable functions than Riemann integrable functions. There is no difference between proper and improper integrals. More precisely, the K-integral  $\int_a^b f(x) dx$  exists if and only if the one-sided limit

$$\lim_{t \rightarrow a-} \int_t^b f(x) dx$$

is well defined and their values coincide, and similarly for the upper limit  $b$ . This is due to the freedom in the choice of the gauges.

There is only an indirect proof that there are bounded functions on a compact interval which are not K-integrable, based on some set-theoretic arguments, but there are no explicit constructions of such functions available.

We say that a set of real numbers  $M$  is measurable if the K-integral of its indicator function  $\chi_M$  exists. The assignment  $m : M \mapsto \int_a^b \chi_M(x) dx$  for all sets  $M \subset [a, b]$  has the properties of a *measure*. The set of such measurable sets  $M$  is closed under finite intersections and countable unions. The measure  $m$  is additive with respect to unions of at most countable systems of pairwise disjoint sets.

This measure coincides with the *Lebesgue measure*. This measure is used in another concept of integration, which is extremely useful in higher dimensional applications, the *Lebesgue integral*. We do not go into more details here. We remark that a real function  $f$  is Lebesgue integrable if and only if its absolute value is K-integrable.

A big advantage of the K-integral compared to other concepts is the possibility of integrating many functions which are not integrable in absolute value. Compare the concepts of convergence and absolute convergence of series.

A typical example is the sinus integral over all reals. The K-integral of the sinc function  $\int_0^\infty \frac{\sin x}{x} dx$  exists, while the absolute value  $g(x) = |\text{sinc}(x)|$  is not Lebesgue integrable. Such integrals are important in models for signal processing where it is necessary to aggregate potentially infinite many interferences canceling each other by different signs.

**6.3.19. The convergence theorems.** We have dealt with uniform convergence and Riemann integrability. With the K-integral, there is a much nicer and stronger theorem available. A special case is the monotone convergence theorem for uniformly bounded functions  $f_0(x) \leq f_1(x) \leq \dots$ .

#### DOMINATED CONVERGENCE THEOREM

**Theorem.** *Suppose  $f_0, f_1, f_2, \dots$  are K-integrable functions on an interval  $[a, b]$ , converging pointwise to the limit function  $f$ . If there are two K-integrable functions  $g$  and  $h$  satisfying*

$$g(x) \leq f_n \leq h(x),$$

*for all  $n \in \mathbb{N}$  and  $x \in [a, b]$ , then  $f$  is K-integrable too, and*

$$\int_a^b f(t) dt = \lim_{n \rightarrow \infty} \int_a^b f_n(t) dt.$$

For monotone convergence, there is a stronger result saying that a sufficient and necessary condition for the K-integrability of the pointwise limit is  $\sup_n \int_a^b f_n(x) dx < \infty$ .

This theorem could not be applied in our third example in 6.3.2. There the functions  $f_n$  have a "bump" which gets larger but narrower when close to the origin. The functions cannot be dominated by an integrable function.

With the Riemann integral, a similar dominated convergence theorem can be proved, except that we have to guarantee the integrability of the pointwise limit  $f$ .

**D. Additional exercises for the whole chapter**



Next we will describe some extra material on the notions that we have analyzed so far in Chapter 6. We begin with additional exercises on higher-order derivatives, Taylor polynomials, Taylor series and other tasks related to differentiation. Some of them are technical, and some other require a bit of our imagination.

**A) MATERIAL ON DERIVATIVES OF HIGHER ORDERS**

**6.D.1.** Prove that  $n$ th derivative of the function  $f(x) = \frac{1}{x+1}$  with  $x \in A = \mathbb{R} \setminus \{-1\}$ , is given by  $f^{(n)}(x) = \frac{(-1)^n n!}{(x+1)^{n+1}}$  for all  $x \in A$ .

**Solution.** For any  $x$  in the domain of  $f$  we have  $f'(x) = -\frac{1}{(x+1)^2}$ . Thus for  $n = 1$  the given relation holds. Assume that  $f^{(k)}(x) = \frac{(-1)^k k!}{(x+1)^{k+1}}$  for some positive integer  $k$ . Then we see that

$$\begin{aligned} f^{(k+1)}(x) &= \left( \frac{(-1)^k k!}{(x+1)^{k+1}} \right)' = (-1)^k k! \left( \frac{1}{(x+1)^{k+1}} \right)' = (-1)^k k! \left( \frac{-(k+1)(x+1)^k (x+1)'}{(x+1)^{2(k+1)}} \right) \\ &= \frac{(-1)^{k+1} (k+1)!}{(x+1)^{k+2}}, \end{aligned}$$

for all  $x \in A$ , where we used the identity  $k!(k+1) = (k+1)!$ . This shows that  $f^{(k+1)}(x) = \frac{(-1)^{k+1} (k+1)!}{(x+1)^{k+2}}$  and the claim follows by the principle of induction.  $\square$

**6.D.2.** Compute the  $n$ th derivative of the function  $f(x) = \frac{1}{x(1-x)}$  with  $x \in \mathbb{R} \setminus \{0, 1\}$ .

**Solution.** For all  $x$  in the domain of  $f$  one can write  $f(x) = g(x) + h(x)$ , where  $g(x) = \frac{1}{x}$  and  $h(x) = \frac{1}{1-x}$ , respectively. Thus it follows that  $f^{(n)}(x) = g^{(n)}(x) + h^{(n)}(x)$  for all  $x$ , and it suffices to compute the  $n$ th derivatives of the functions  $g, h$ . For  $g$  we see that

$$g'(x) = -x^{-2}, \quad g''(x) = 2x^{-3}, \quad g'''(x) = -6x^{-4} = -1 \cdot 2 \cdot 3x^{-4}, \quad g^{(4)}(x) = 24x^{-5} = 1 \cdot 2 \cdot 3 \cdot 4x^{-5}.$$

Hence clearly we have

$$g^{(n)}(x) = (-1)^n n! x^{-(n+1)} = \frac{(-1)^n n!}{x^{n+1}},$$

which easily follows by induction. Similarly, for  $h(x)$  we can prove that  $h^{(n)}(x) = \frac{n!}{(1-x)^{n+1}}$ . Thus finally we deduce that

$$f^{(n)}(x) = n! \left( \frac{(-1)^n}{x^{n+1}} + \frac{1}{(1-x)^{n+1}} \right), \quad x \in \mathbb{R} \setminus \{0, 1\}.$$

$\square$

**6.D.3.** Consider the function  $f(x) = 1/x$ , with  $x \in \mathbb{R} \setminus \{0\}$ , and let  $n$  be a positive integer. Compute the limit  $\lim_{x \rightarrow +\infty} |f^{(n)}(x)|$ , for all  $x \neq 0$ . Next confirm your answer via Sage by combining the commands `limit` and `assume`.

**Solution.** First we need to compute the  $n$ th derivative. We see that

$$f'(x) = -x^{-2}, \quad f''(x) = 2x^{-3}, \quad f'''(x) = -2 \cdot 3x^{-4}, \quad f^{(4)}(x) = 2 \cdot 3 \cdot 4x^{-5}, \quad \dots, \quad f^{(n)}(x) = (-1)^n n! x^{-n-1}.$$

Hence  $f^{(n)}(x) = (-1)^n n! x^{-n-1}$  and we leave as practice the proof by induction over  $n$ . To compute the given limit one may use the ratio test:

$$\lim_{n \rightarrow +\infty} \left| \frac{f^{(n+1)}(x)}{f^{(n)}(x)} \right| = \lim_{n \rightarrow +\infty} \left| \frac{(-1)^{n+1} (n+1)! x^{-n-2}}{(-1)^n n! x^{-n-1}} \right| = \lim_{n \rightarrow +\infty} \frac{(n+1)}{|x|} = +\infty.$$

As for a Sage verification, it is necessary to use the command `assume` as follows

```
var("n"); assume(x>1)
limit(abs((-1)^n*(factorial(n))*(x**(-n-1))), n=oo)
```

Otherwise, Sage prints out an error, suggesting the use of further restrictions.  $\square$

**6.D.4.** Determine the Taylor expansion of third order around the point  $a = 0$  of the following functions:

$$g(x) = \frac{1}{\cos(x)}, \quad h(x) = e^{-\frac{x^2}{2}}, \quad k(x) = \sin(\sin(x)).$$

Next verify your answers in Sage.  $\circ$

**6.D.5.** For the function  $f(x) = e^{\sin(x)}$  with  $x \in \mathbb{R}$  compute the Taylor polynomials  $T_a^k f(x)$ , where  $k = 1, 2, 3, 4$  and  $a = 0$ . Deduce that  $T_0^3 f(x) = T_0^2 f(x)$  for all  $x$ . Moreover, use Sage to plot the graphs of these Taylor extensions together with the graph of  $f$ .

**Solution.** We have  $f(0) = e^0 = 1$ , and compute

$$f'(0) = \cos(x) e^{\sin(x)} \Big|_{x=0} = 1, \quad f''(0) = \left[ e^{\sin(x)} (\cos^2(x) - \sin(x)) \right] \Big|_{x=0} = 1,$$

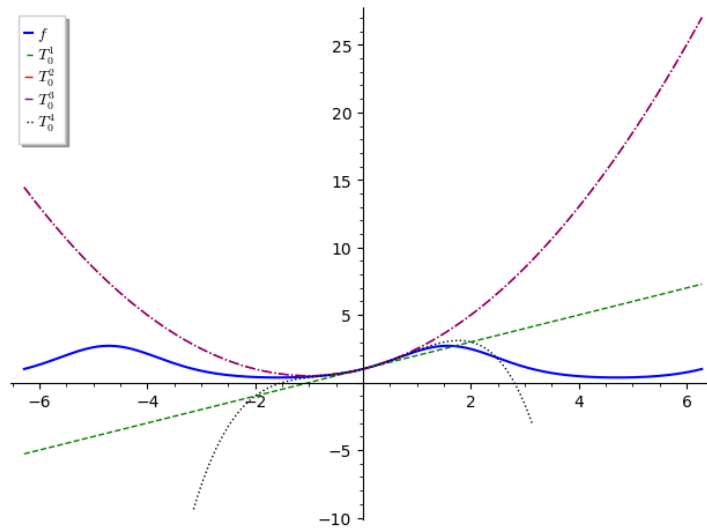
$$f'''(0) = \left[ e^{\sin(x)} (\cos^3(x) - 3 \cos(x) \sin(x) - \cos(x)) \right] \Big|_{x=0} = 1 - 1 = 0,$$

$$f^{(4)}(0) = \left[ e^{\sin(x)} (\cos^4(x) - 6 \cos^2(x) \sin(x) - 4 \cos^2(x) + 3 \sin^2(x) + \sin(x)) \right] \Big|_{x=0} = 1 - 4 = -3.$$

Thus

$$T_0^1 f(x) = 1 + x, \quad T_0^2 f(x) = 1 + x + \frac{1}{2}x^2, \quad T_0^3 f(x) = T_0^2 f(x), \quad T_0^4 f(x) = 1 + x + \frac{1}{2}x^2 - \frac{1}{8}x^4.$$

In the figure below we present the graphs of these polynomials.



To produce this figure in Sage (together with a confirmation of the expressions of  $T_0^k f(x)$  for  $k = 1, \dots, 4$ ) we applied the following block:

```
f(x)=exp(sin(x))
p=plot(f(x), x, -2*pi, 2*pi, thickness=1.5, legend_label=r"$f$")
T1(x)=taylor(f(x), x, 0, 1)
p1=plot(T1(x), x, -2*pi, 2*pi, color="green", linestyle="--", legend_label=r"$T^1_{0}$")
T2(x)=taylor(f(x), x, 0, 2)
p2=plot(T2(x), x, -2*pi, 2*pi, color="red", linestyle="-.", legend_label=r"$T^2_{0}$")
T3(x)=taylor(f(x), x, 0, 3)
p3=plot(T3(x), x, -2*pi, 2*pi, color="purple", linestyle="-.", legend_label=r"$T^3_{0}$")
T4(x)=taylor(f(x), x, 0, 4)
p4=plot(T4(x), x, -pi, pi, color="black", linestyle=":", legend_label=r"$T^4_{0}$")
show(p+p1+p2+p3+p4)
```

□

**6.D.6.** Consider the function  $f(x) = \ln(1+x^2)/(1-\cos(x))$ . Using the Taylor series centered at 0 of the numerator and dominator of  $f$ , compute  $\lim_{x \rightarrow 0} f(x)$ . Next present a confirmation of your result based on the l'Hopital's rule.

**Solution.** Recall that the Taylor series centered at 0 of  $\ln(1+x)$  and  $\cos(x)$  are respectively given by

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} x^n}{n}, \quad |x| < 1, \quad \cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} \quad x \in \mathbb{R}.$$

Thus, the Taylor series around 0 of  $\ln(1+x^2)$  has the form

$$\ln(1+x^2) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} x^{2n} = x^2 - \frac{x^4}{2} + \frac{x^6}{3} - \frac{x^8}{4} + \dots, \quad |x| < 1,$$

while

$$1 - \cos(x) = 1 - \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots\right) = \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} - \frac{x^8}{8!} + \dots, \quad x \in \mathbb{R}.$$

We can now use these two formulas to compute the limit at hand:

$$\begin{aligned} \lim_{x \rightarrow 0} f(x) &= \lim_{x \rightarrow 0} \frac{\ln(1+x^2)}{1-\cos(x)} = \lim_{x \rightarrow 0} \frac{x^2 - \frac{x^4}{2} + \frac{x^6}{3} - \frac{x^8}{4} + \dots}{\frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} - \frac{x^8}{8!} + \dots} = \lim_{x \rightarrow 0} \frac{x^2 \left(1 - \frac{x^2}{2} + \frac{x^4}{3} - \frac{x^6}{4} + \dots\right)}{x^2 \left(\frac{1}{2} - \frac{x^2}{4!} + \frac{x^4}{6!} - \frac{x^6}{8!} + \dots\right)} \\ &= \lim_{x \rightarrow 0} \frac{1 - \frac{x^2}{2} + \frac{x^4}{3} - \frac{x^6}{4} + \dots}{\frac{1}{2} - \frac{x^2}{4!} + \frac{x^4}{6!} - \frac{x^6}{8!} + \dots} = \frac{1}{\frac{1}{2}} = 2. \end{aligned}$$

To confirm this result by an alternative method, apply the l'Hopital rule twice:

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{\ln(1+x^2)}{1-\cos(x)} &\stackrel{0}{=} \lim_{x \rightarrow 0} \frac{(\ln(1+x^2))'}{(1-\cos(x))'} = \lim_{x \rightarrow 0} \frac{\frac{2x}{1+x^2}}{\sin(x)} = \lim_{x \rightarrow 0} \frac{2x}{(1+x^2)\sin(x)} \\ &\stackrel{0}{=} \lim_{x \rightarrow 0} \frac{(2x)'}{((1+x^2)\sin(x))'} = \lim_{x \rightarrow 0} \frac{2}{2x\sin(x) + (1+x^2)\cos(x)} = \frac{2}{0+1} = 2. \end{aligned}$$

□

**6.D.7.** Consider the function  $f(x) = \frac{e^x - e^{-x} - 2x}{x^2 - x \ln(x+1)}$ . Using appropriately the theory of Taylor series to evaluate the limit  $\lim_{x \rightarrow 0} f(x)$ . Next confirm your computation by Sage.

**Solution.** Observe that limit that we need to compute has the indeterminate form 0/0. We have

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} + \dots, \quad e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots + \frac{(-1)^n x^n}{n!} + \dots$$

Thus,  $e^x - e^{-x} = 2x + \frac{2x^3}{3!} +$  higher order terms. This implies that the numerator  $e^x - e^{-x} - 2x$  is dominated by the term  $\frac{2x^3}{3!}$ , as  $x \rightarrow 0$ . Similarly,

$$x \ln(x+1) = x \left(x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots\right) = x^2 - \frac{x^3}{2} + \frac{x^4}{3} - \frac{x^5}{4} + \dots.$$

Thus the dominator  $x^2 - x \ln(1+x)$  is dominated by the term  $\frac{x^3}{2}$ , as  $x \rightarrow 0$ . Combining these two observations we get  $\lim_{x \rightarrow 0} f(x) = 2/3$ .

In Sage to confirm the result give the cell

```
f(x) = (e^x - e^(-x) - 2*x) / (x^2 - x*ln(1+x)); limit(f(x), x=0)
```

□

**6.D.8.** Consider the function

$$f(x) = \begin{cases} 1/x^2, & \text{if } x \neq 0, \\ 0, & \text{if } x = 0. \end{cases}$$

Determine the intervals of monotonicity and find the extremes points of  $f$ , if any.

**Solution.** The domain of  $f$  is the whole real line, but  $f$  is not continuous at  $x = 0$ , since

$$\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^-} f(x) = \infty,$$

and  $f(0) = 0$  (try to plot the graph of  $f$ ). Recall in Sage we can get the same, just by the cell

```
show(lim(1/x^2, x=0, dir="left")); show(lim(1/x^2, x=0, dir="right"))
```

Hence,  $f$  is neither differentiable at  $x = 0$ , i.e.,  $f'(0)$  does not exist. Now, for all  $x \neq 0$  we have  $f'(x) = -2/x^3$  and hence  $f$  is strictly increasing for  $x < 0$  and strictly decreasing for  $x > 0$ . However, at  $x = 0$  the function  $f$  does not a maximum, since it is discontinuous at this point. □

**6.D.9.** Find the asymptotes of the following functions:



$$(a) y = 2 \arctan \left( \left| \frac{x}{x^2 - 1} \right| \right), \quad x \in \mathbb{R} \setminus \{\pm 1\}, \quad (b) y = \ln \left( \frac{3e^{2x} + e^x + 10}{e^x + 1} \right), \quad x \in \mathbb{R}.$$

○

**6.D.10.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a differentiable function satisfying  $f(x) + e^{f(x)} = x + 1$ , for all  $x \in \mathbb{R}$ .

(a) Show that  $f(0) = 0$ .

(b) Does  $f$  have some critical points? Show that  $f$  is strictly increasing.

(c) Show that  $f$  is positive for  $x > 0$  and negative for  $x < 0$ .

(d) Prove the inequality  $\frac{x}{2} \leq f(x) \leq xf'(x)$ , for all  $x \in \mathbb{R}$ .

**Solution.** (a) The defining equation  $f(x) + e^{f(x)} = x + 1$  for  $x = 0$  gives

$$f(0) + e^{f(0)} - 1 = 0. \quad (*)$$

Suppose that  $f(0) = y \in \mathbb{R}$  and consider the function  $g(y) = y + e^y - 1$ . Obviously,  $y = 0$  is root of  $g$ , i.e.,  $g(0) = 0$ . On the other hand, the first derivative of  $g$  is given by  $g'(y) = 1 + e^y$  and we have  $g'(y) > 0$  for all  $y \in \mathbb{R}$ . Thus the function  $g$  is strictly increasing, which implies that  $y = 0$  is its unique root. Thus  $f(0) = 0$ .

Another solution: For  $y = f(0)$  and  $g(y) = y + e^y - 1$  the relation (\*) is written as  $g(f(0)) = 0$ . However,  $g$  is injective since  $g$  is strictly increasing. Therefore, the relation  $g(f(0)) = 0$  gives  $f(0) = 0$ .

(b) By differentiating the relation  $f(x) + e^{f(x)} = x + 1$  with respect to  $x$  we get  $f'(x)(1 + e^{f(x)}) = 1$ . Thus

$$f'(x) = \frac{1}{1 + e^{f(x)}} > 0, \quad x \in \mathbb{R}. \quad (**)$$

Hence  $f$  has no critical points, and in particular is strictly increasing.

(c) To obtain this claim one can rely on the monotonicity of  $f$  and use (a). In particular, we have

$$x > 0 \Leftrightarrow f(x) > f(0) = 0, \quad x < 0 \Leftrightarrow f(x) < f(0) = 0.$$

(d) Here we will use the second derivative of  $f$ , given by

$$f''(x) = \left( \frac{1}{1 + e^{f(x)}} \right)' = -\frac{f'(x) e^{f(x)}}{(1 + e^{f(x)})^2} = -\frac{e^{f(x)}}{(1 + e^{f(x)})^3} < 0,$$

for all  $x \in \mathbb{R}$ . Notice also by (\*\*) that  $f'(0) = 1/2$ . Now, for  $x > 0$  the given inequality holds as an equality. For  $x > 0$  consider the set  $[0, x]$ . On this set  $f$  satisfies the conditions of the mean value theorem, hence there exist  $\xi \in (0, x)$  such that

$$f'(\xi) = \frac{f(x) - f(0)}{x - 0} = \frac{f(x)}{x}, \quad (\#)$$

where we used (a) to replace  $f(0)$  by 0. On the other hand we saw that  $f'$  is strictly decreasing (since  $f''(x) < 0$  for all  $x$ ). Thus, in combination with (#) we obtain the following equivalences:

$$0 < \xi < x \Leftrightarrow f'(x) > f'(\xi) > f'(0) \Leftrightarrow f'(x) > \frac{f(x)}{x} > \frac{1}{2} \Leftrightarrow xf'(x) > f(x) > \frac{x}{2}$$

which proves the given formula for  $x > 0$ . For  $x < 0$  one proceeds similarly by applying the mean value theorem on the set  $[x, 0]$ . □

**6.D.11.** Study the local behaviour of the function  $f(x) = \sqrt[3]{|x|^3 + 1}$ , with  $x \in \mathbb{R}$ .

**Solution.** It is easy to deduce that  $f$  is continuous everywhere on  $\mathbb{R}$ . Also,  $f(x) \geq 1$  and  $f(-x) = f(x)$  for all  $x \in \mathbb{R}$ , i.e., the function  $f$  is positive and even. As for the limit behavior of  $f$  at  $\pm\infty$ , we compute (see also below the graph of  $f$ )

$$\lim_{x \rightarrow \pm\infty} \sqrt[3]{|x|^3 + 1} = \lim_{x \rightarrow \pm\infty} \sqrt[3]{|x|^3} = \lim_{x \rightarrow \pm\infty} |x| = +\infty.$$

For the first derivative of  $f$  we compute

$$f'(x) = \begin{cases} \frac{x^2}{\sqrt[3]{(x^3+1)^2}}, & \text{for } x > 0, \\ 0, & \text{for } x = 0, \\ -\frac{x^2}{\sqrt[3]{(-x^3+1)^2}}, & \text{for } x < 0. \end{cases}$$

Notice for computing  $f'(0)$  we used the one-side limits

$$\lim_{x \rightarrow 0^+} \frac{x^2}{\sqrt[3]{(x^3+1)^2}} = 0 = \lim_{x \rightarrow 0^-} -\frac{x^2}{\sqrt[3]{(-x^3+1)^2}}.$$

We deduce that  $f'(x) > 0$  for all  $x > 0$  and  $f'(x) < 0$  for all  $x < 0$ . This implies that  $f$  is strictly increasing on the interval  $(0, +\infty)$ , increasing on  $[0, +\infty)$ , strictly decreasing for all  $x \in (-\infty, 0)$ , and increasing on the interval  $(-\infty, 0]$  (recall that  $f$  is an even function). It follows that  $f$  has only one local minimum at point  $x_0 = 0$ , which is a global minimum with value  $f(0) = 1$ .

Next we will prove that  $f$  is convex for any  $x \in \mathbb{R}$ . Indeed, the second derivative of  $f$  is given by

$$f''(x) = \begin{cases} \frac{2x}{\sqrt[3]{(x^3+1)^5}}, & \text{if } x > 0; \\ 0, & \text{if } x = 0; \\ -\frac{2x}{\sqrt[3]{(-x^3+1)^5}}, & \text{if } x < 0. \end{cases}$$

Thus we deduce that  $f''(x) > 0$  for all  $x \in (-\infty, 0) \cup (0, \infty)$ , so  $f$  is strictly convex on this interval. In total, we obtained that  $f$  is convex on its whole domain (it doesn't have some inflection points).

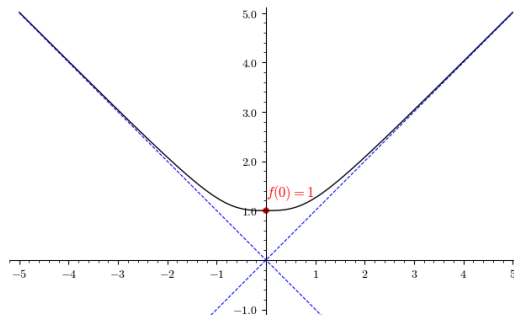
Let us finally determine the asymptotes of  $C_f$ . Recall that a line  $y = ax + b$  is an inclined asymptote for  $x \rightarrow \infty$ , if and only if both (proper) limits

$$\lim_{x \rightarrow \infty} \frac{f(x)}{x} = a, \quad \lim_{x \rightarrow \infty} (f(x) - ax) = b,$$

exist. Analogous statement holds for  $x \rightarrow -\infty$ . Hence the limits

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{f(x)}{x} &= \lim_{x \rightarrow \infty} \frac{\sqrt[3]{x^3+1}}{x} = \lim_{x \rightarrow \infty} \frac{\sqrt[3]{x^3}}{x} = 1, \\ \lim_{x \rightarrow \infty} (f(x) - 1 \cdot x) &= \lim_{x \rightarrow \infty} \left( \sqrt[3]{x^3+1} - x \right) = \lim_{x \rightarrow \infty} \left( \left[ \sqrt[3]{x^3+1} - x \right] \frac{\sqrt[3]{(x^3+1)^2} + x\sqrt[3]{x^3+1} + x^2}{\sqrt[3]{(x^3+1)^2} + x\sqrt[3]{x^3+1} + x^2} \right) \\ &= \lim_{x \rightarrow \infty} \frac{x^3+1-x^3}{\sqrt[3]{(x^3+1)^2} + x\sqrt[3]{x^3+1} + x^2} = \lim_{x \rightarrow \infty} \frac{1}{3x^2} = 0, \end{aligned}$$

imply that the line  $y = x$  is an asymptote at  $+\infty$ . Since  $f$  is even, we immediately obtain the line  $y = -x$  as an asymptote at  $-\infty$ . Let us finally present the graph of  $f$  together with the asymptotes  $y = \pm x$ :



□

**6.D.12.** Using Sage as a tool, examine the local behaviour of the function  $f(x) = \arctan\left(\frac{x}{2-x}\right)$ .

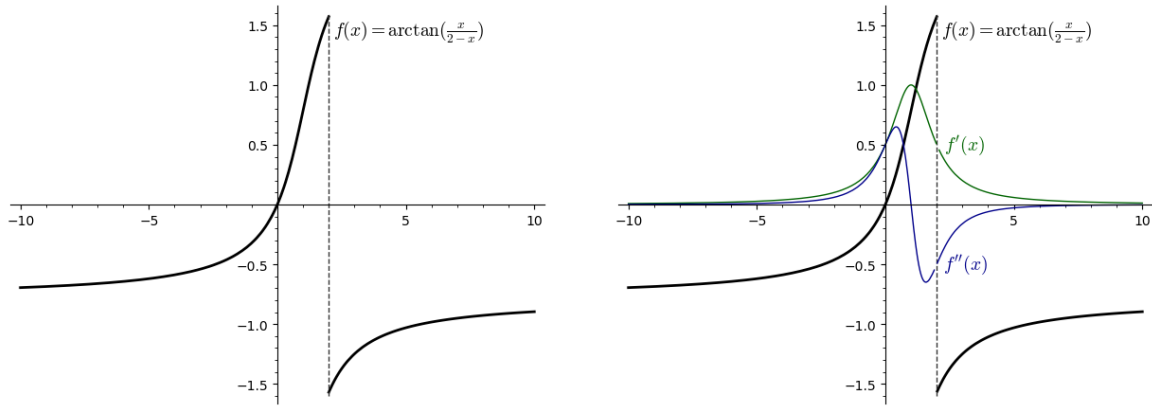
**Solution.** The given function is defined on  $A = \mathbb{R} \setminus \{2\}$ . Its graph does not have special symmetry. The only point of intersection of  $C_f$  with the axes is the origin  $[0, 0]$ , while  $f$  is positive exactly on the open interval  $(0, 2)$ . In Sage the roots of  $f$  occur by the command

```
f(x)=arctan(x/(2-x)); show(solve(f(x)==0, x))
```

while one can verify that  $f$  is not even/odd by adding the cell

```
show(bool(f(-x)==f(x))); show(bool(f(-x)==-f(x)))
```

where for both cases Sage returns False. The function is everywhere continuous on  $A$ , and at  $x_0 = 2$  the graph of  $f$  has a jump of size  $\pi$ . This follows from the limits  $\lim_{x \rightarrow 2^-} f(x) = \frac{\pi}{2}$  and  $\lim_{x \rightarrow 2^+} f(x) = -\frac{\pi}{2}$ , see also the graph  $C_f$  of  $f$  below:



This illustration occurs by adding in the initial block the cell

```
p=plot(f(x), x, -10, 10, exclude=[2], color="black", thickness=2)
p+=text(r"$f(x)=\arctan(\frac{x}{2-x})$", (5.1, 1.45), fontsize=14, color="black")
p+=line([(2, -1.6), (2, 1.6)], rgbcolor=(0.2,0.2,0.2), linestyle="--")
p.show(figsize=6)
```

Moreover we see that  $\lim_{x \rightarrow -\infty} f(x) = -\frac{\pi}{4} = \lim_{x \rightarrow +\infty} f(x)$ , and thus the range of  $f$  is the set  $(-\pi/2, \pi/2) \setminus \{-\pi/4\}$ . Moreover, the only asymptote of  $C_f$  is the line  $y = -\pi/4$  at  $\pm\infty$ . We also compute

$$f'(x) = \frac{1}{x^2 - 2x + 2}, \quad f''(x) = \frac{2(1-x)}{(x^2 - 2x + 2)^2}, \quad x \in \mathbb{R} \setminus \{2\}.$$

The limits posed above together with the first two derivatives can be confirmed by adding in the initial block the cell

```
show(lim(f(x), x=2, dir="+")); show(lim(f(x), x=2, dir="-"))
show(lim(f(x), x=oo,)); show(lim(f(x), x=-oo))
show(diff(f, x).factor()); show(diff(f, x, 2).factor())
```

It is easy to see that  $f'(x) > 0$  for all  $x \in A$  and hence  $f$  is increasing at every point of its domain, see also the graph of  $f'(x)$  above (which is coloured by green). This claim can be confirmed in Sage by adding the cell

```
A=RealSet(x!=2); bool(diff(f, x)>0 for x in A)
```

Notice the first command in this cell declares the domain  $A$  of  $f$ . Finally, the function  $f$  is convex on the interval  $(-\infty, 1)$ , and concave on the interval  $(1, 2)$ ,  $(2, +\infty)$ . In the figure at the right we included the graph of  $f''(x)$ , which provides a graphical proof of these claims. The point  $x_1 = 1$  is the unique point of inflection with  $f(1) = \pi/4$ , which we can easily confirm by adding the cell `solve(diff(f, x, 2) == 0, x)`.  $\square$

**6.D.13.** Study the local behaviour of the function  $f(x) = -\frac{x^2}{x+1}$ , with  $x \in \mathbb{R} \setminus \{-1\}$  and use Sage to sketch the graph of  $f$  together with its asymptotes. ○

**6.D.14.** Study the local behaviour of the function  $f(x) = \frac{x^3 - 3x^2 + 3x + 1}{x - 1}$  and use Sage to confirm most of your computations. ○

**6.D.15.** Study the local behaviour of the function  $f(x) = \sqrt[3]{x}e^{-x}$ . ○

Consider a set of  $n$  functions  $\{f_1(t), \dots, f_n(t)\}$  which are  $(n - 1)$  times differentiable. The Wronski matrix associated to this set is defined by

$$W(f_1, \dots, f_n) = \begin{pmatrix} f_1(t) & f_2(t) & \dots & f_n(t) \\ f_1'(t) & f_2'(t) & \dots & f_n'(t) \\ f_1''(t) & f_2''(t) & \dots & f_n''(t) \\ \vdots & \vdots & \vdots & \vdots \\ f_1^{(n-1)}(t) & f_2^{(n-1)}(t) & \dots & f_n^{(n-1)}(t) \end{pmatrix}.$$

One can prove that the set of functions  $\{f_1(t), \dots, f_n(t)\}$  is linearly independent if and only if  $\det(W) \neq 0$ . Let us describe an application of this result, which relates linear algebra with calculus.

**6.D.16. Wronki matrix.** Consider the following set  $B = \{\sin(t), \cos(t), e^t\}$ .

(a) Show that  $B$  consists of linearly independent functions.

(b) If  $V$  is the real vector space of functions spanned by  $B$ , let  $\mathcal{D} : V \rightarrow V$  be the linear operator defined by  $\mathcal{D}(f) = f'(t) = \frac{d f(t)}{dt}$ . Compute the matrix of  $\mathcal{D}$  with respect to the basis  $B$ . Moreover, find its kernel and its image.

**Solution.** For the set  $B = \{\sin(t), \cos(t), e^t\}$  we compute

$$\det(W(\sin, \cos, \exp)) = \begin{vmatrix} \sin(t) & \cos(t) & e^t \\ \cos(t) & -\sin(t) & e^t \\ -\sin(t) & -\cos(t) & e^t \end{vmatrix} = -2e^t \neq 0,$$

hence  $B$  consists of linearly independent vectors. We can easily verify the computation of the determinant in Sage, by typing

```
t = var("t")
W = matrix(SR, 3, 3, [[sin(t), cos(t), exp(t)], [cos(t), -sin(t), exp(t)], [-sin(t), -cos(t), exp(t)]])
W; w=W.det (); w.full_simplify ()
```

which returns the desired answer  $-2 * e^t$ .

(b) For the matrix of the linear operator  $\mathcal{D}$  with respect to the basis  $B$  we see that

$$\mathcal{D}(\sin(t)) = 0 \sin(t) + 1 \cos(t) + 0e^t, \quad \mathcal{D}(\cos(t)) = -1 \sin(t) + 0 \cos(t) + 0e^t, \quad \mathcal{D}(e^t) = 0 \sin(t) + 0 \cos(t) + 1e^t.$$

Then the coordinates of  $\mathcal{D}(\sin(t))$ ,  $\mathcal{D}(\cos(t))$  and  $\mathcal{D}(e^t)$  give the columns of the matrix of  $\mathcal{D}$ , that is,  $\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

Let us now quickly compute the kernel of  $\mathcal{D}$  via Sage:

```
D = matrix(SR, 3, 3, [[0, -1, 0], [1, 0, 0], [0, 0, 1]])
D.right_kernel ()
```

which prints out

```
Vector space of degree 3 and dimension 0 over Symbolic Ring
Basis matrix:
[]
```

This essentially means that  $\text{Ker}(\mathcal{D}) = \{0\}$ . For the image use the command `D.image ()`. It gives<sup>5</sup>

```
Vector space of degree 3 and dimension 3 over Symbolic Ring
Basis matrix:
[1 0 0]
[0 1 0]
[0 0 1]
```

Thus the image of  $\mathcal{D}$  is spanned by the standard basis of  $\mathbb{R}^3$ . □

*We proceed with material on parametrized curves and surfaces.*

**6.D.17.** Find the points on the cycloid given in 6.A.45 where the tangent lines are vertical/parallel. ○

*In differential calculus there is a plethora of important inequalities. Among them there is one with many applications, the so called Jensen inequality, This is about convex and concave functions, as we will see below.*

**6.D.18. Jensen inequality.** For a strictly convex function  $f$  on interval  $I$  and for arbitrary points  $x_1, \dots, x_n \in I$  and real numbers  $c_1, \dots, c_n > 0$  such that  $c_1 + \dots + c_n = 1$ , the inequality

$$f\left(\sum_{i=1}^n c_i x_i\right) \leq \sum_{i=1}^n c_i f(x_i)$$

holds, with equality occurring if and only if  $x_1 = \dots = x_n$ . Notice the Jensen inequality can be also formulated in a more intuitive way: “the centroid of mass points placed upon a graph of a strictly convex function lies above this graph.”

**Solution.** Could be proven easily by induction: for  $n = 2$  it is just the definition of the convex function, for the induction step

$$f\left(\sum_{i=1}^{k+1} c_i x_i\right) = f\left(c_1 x_1 + (1 - c_1) \sum_{i=2}^{k+1} \frac{c_i}{1 - c_1} x_i\right) \leq c_1 f(x_1) + (1 - c_1) f\left(\sum_{i=2}^{k+1} \frac{c_i}{1 - c_1} x_i\right)$$

<sup>5</sup>Recall that the basis computed by Sage is “row reduced”.

$$\leq c_1 f(x_1) + (1 - c_1) \left( \sum_{i=2}^{k+1} \frac{c_i}{1 - c_1} f(x_i) \right) = \sum_{i=1}^{k+1} c_i f(x_i).$$

Notice above we used the inequality first for  $n = 2$  and then for  $n = k$ .  $\square$

**6.D.19.** Prove that among all (convex)  $n$ -gons inscribed into a circle, the regular  $n$ -gon has the largest area (for arbitrary  $n \geq 3$ ).

**Solution.** Clearly it suffices to consider the  $n$ -gons inside of which lies the center of the circle. We'll divide each such  $n$ -gon inscribed into a circle with radius  $r$  to  $n$  triangles with areas  $S_i, i \in \{1, \dots, n\}$  according to the figure. With regard to the fact that

$$\sin \frac{\varphi_i}{2} = \frac{x_i}{r}, \quad \cos \frac{\varphi_i}{2} = \frac{h_i}{r}, \quad i \in \{1, \dots, n\},$$

we have

$$S_i = x_i h_i = r^2 \sin \frac{\varphi_i}{2} \cos \frac{\varphi_i}{2} = \frac{1}{2} r^2 \sin \varphi_i, \quad i \in \{1, \dots, n\}.$$

This implies that the area of the hole  $n$ -gon is

$$S = \sum_{i=1}^n S_i = \frac{1}{2} r^2 \sum_{i=1}^n \sin \varphi_i.$$

Thus we want to maximize the sum  $\sum_{i=1}^n \sin \varphi_i$ , while for values  $\varphi_i \in (0, \pi)$  we clearly have

$$(1) \quad \varphi_1 + \dots + \varphi_n = \sum_{i=1}^n \varphi_i = 2\pi.$$

The function  $y = \sin x$  is strictly concave on the interval  $(0, \pi)$ , which means, that the function  $y = -\sin x$  is strictly convex on this interval. Then according to Jensen's inequality for  $c_i = 1/n$  and  $x_i = \varphi_i$ , we have

$$-\sin \left( \sum_{i=1}^n \frac{1}{n} \varphi_i \right) \leq -\sum_{i=1}^n \frac{1}{n} \sin \varphi_i, \quad \text{tj.} \quad \sin \left( \sum_{i=1}^n \frac{1}{n} \varphi_i \right) \geq \sum_{i=1}^n \frac{1}{n} \sin \varphi_i.$$

Moreover, we know the equality occurs exactly for  $\varphi_1 = \dots = \varphi_n$ . If we express (using (1))

$$S = \frac{r^2 n}{2} \sum_{i=1}^n \frac{1}{n} \sin \varphi_i \leq \frac{r^2 n}{2} \sin \left( \sum_{i=1}^n \frac{1}{n} \varphi_i \right) = \frac{r^2 n}{2} \sin \frac{2\pi}{n},$$

we can see that  $S$  can attain at most the value on the right hand side. But that happens if and only if  $\varphi_1 = \dots = \varphi_n$  (we chose  $x_i = \varphi_i$ ). Hence the regular  $n$ -gon is the one with the maximum area, because it satisfies  $\varphi_1 = \dots = \varphi_n = 2\pi/n$ .  $\square$

**6.D.20. Isoperimetric quotient.** For a closed curve in plane enclosing a planar region, we define its isoperimetric quotient as the number

$$IQ := \frac{S}{\pi \left(\frac{o}{2\pi}\right)^2} = \frac{4\pi S}{o^2},$$

where  $S$  denotes the area of the region and  $o$  its perimeter (i.e. the length of the curve). Hence the isoperimetric quotient determines the ratio of the area of the region and the area of a circle with the same perimeter as the given region. The notation  $IQ$  is therefore not only an English abbreviation for the isoperimetric quotient, but can be also thought of as the "intelligence of the region", with which it uses its perimeter for attaining as big area as possible. The isoperimetric theorem then states that for every closed curve,  $IQ \leq 1$ , with equality occurring only for a circle, or ("the circle is the smartest").

Determine  $IQ$  for a regular polygon and a circle and find the sector of a circle, for which its boundary has the largest  $IQ$

**Solution.** First notice that the value of  $IQ$  doesn't change with a change of scale on the axes (same on both). Because when the proportions of the region get  $a$  times bigger (for arbitrary  $a > 0$ ), the perimeter also gets  $a$  times bigger and the area  $a^2$  times (it's a square measure). Hence  $IQ$  doesn't depend on the size of the region, but only on its shape. Thus we can consider a regular  $n$ -gon inscribed into a unit circle. According to the figure,

$$h = \cos \varphi = \cos \frac{\pi}{n}, \quad \frac{x}{2} = \sin \varphi = \sin \frac{\pi}{n},$$

which yields

$$o_n = n \cdot x = 2n \sin \frac{\pi}{n}$$

and

$$S_n = n \cdot \frac{1}{2} hx = n \cos \frac{\pi}{n} \sin \frac{\pi}{n}.$$

Thus for a regular  $n$ -gon, we have

$$IQ = \frac{4\pi n \cos \frac{\pi}{n} \sin \frac{\pi}{n}}{4n^2 \sin^2 \frac{\pi}{n}} = \frac{\pi}{n} \cotg \frac{\pi}{n},$$

which we can verify for example for a square ( $n = 4$ ) with a side of length  $a$ , where

$$IQ = \frac{4\pi a^2}{(4a)^2} = \frac{\pi}{4} = \frac{\pi}{4} \cotg \frac{\pi}{4}.$$

Using the limit transition for  $n \rightarrow \infty$  and the limit

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1,$$

we get the isoperimetric quotient for a circle:

$$IQ = \lim_{n \rightarrow \infty} \frac{\pi}{n} \cotg \frac{\pi}{n} = \lim_{n \rightarrow \infty} \frac{\cos \frac{\pi}{n}}{\sin \frac{\pi}{n}} = \frac{\cos 0}{1} = 1.$$

Of course, for a circle with radius  $r$ , we could have also directly computed

$$IQ = \frac{4\pi S}{o^2} = \frac{4\pi(\pi r^2)}{(2\pi r)^2} = 1.$$

For the boundary  $a$  of sector of a circle with radius  $r$  and central angle  $\varphi \in (0, 2\pi)$ , we have

$$IQ = \frac{4\pi S}{o^2} = \frac{4\pi \frac{\varphi r^2}{2}}{(2r+r\varphi)^2} = \frac{2\pi\varphi}{(2+\varphi)^2}.$$

Hence we're looking for a maximum of the function

$$f(\varphi) := \frac{2\pi\varphi}{(2+\varphi)^2}, \quad \varphi \in (0, 2\pi).$$

By computing

$$f'(\varphi) = 2\pi \frac{(2+\varphi)^2 - 2\varphi(2+\varphi)}{(2+\varphi)^4} = 2\pi \frac{2-\varphi}{(2+\varphi)^3}, \quad \varphi \in (0, 2\pi)$$

we easily obtain that

$$f'(\varphi) > 0, \quad \varphi \in (0, 2), \quad f'(\varphi) < 0, \quad \varphi \in (2, 2\pi).$$

Hence function  $f$  attains its maximal value for  $\varphi_0 = 2$  and for a central angle  $\varphi_0 = 2$  (radians), we get the largest

$$IQ = \frac{2\pi\varphi_0}{(2+\varphi_0)^2} = \frac{\pi}{4}.$$

For the sake of completeness, for a solid in three-dimensional space (more precisely, for the closed surface which is its boundary), we define

$$IQ := \frac{V}{\frac{4\pi}{3} \left(\frac{S}{4\pi}\right)^{\frac{3}{2}}},$$

where  $V$  is the volume and  $S$  the surface of the solid. Thus we compare the volume of the solid with a given surface with the volume of the ball with the same space.  $\square$

**6.D.21.** A string of length  $l$  is given. The task is to cut it into  $n$  parts so that it's possible to create boundaries of geometric figures given in advance (for example a square, a triangle, a circle, a halfcircle) with the least sum of areas from the  $n$  smaller strings.



**Solution.** To solve this problem, we'll use the isoperimetric quotient of curves and Jensen's inequality (stated in previous examples). For the geometric figures given in advance, denote the values of their isoperimetric quotients as

$$\frac{1}{\lambda_i} := \frac{4\pi S_i}{o_i^2}, \quad i \in \{1, \dots, n\},$$

where  $S_i$  is the area and  $o_i$  the perimeter of the  $i$ -th figure. We'll also use the denotation

$$\Lambda := \sum_{i=1}^n \lambda_i.$$

Recall that the isoperimetric quotient is given only by the shape of the figure and doesn't depend on its size. In particular, the value  $\Lambda$  is constant (it's determined by the shapes of the given figures).

Our task is to minimize the sum  $\sum_{i=1}^n S_i$  with  $\sum_{i=1}^n o_i = l$ . Because

$$S_i = \frac{o_i^2}{4\pi\lambda_i}, \quad i \in \{1, \dots, n\},$$

we need to minimize the expression

$$S := \frac{1}{4\pi} \sum_{i=1}^n \frac{o_i^2}{\lambda_i}.$$

Using Jensen's inequality for the strictly convex function  $y = x^2$  (on the whole real axis), we obtain

$$\left( \sum_{i=1}^n c_i x_i \right)^2 \leq \sum_{i=1}^n c_i x_i^2$$

for  $x_i \in \mathbb{R}$  and  $c_i > 0$  with the property  $c_1 + \dots + c_n = 1$ . Moreover we know that the equality occurs if and only if  $x_1 = \dots = x_n$ . By choosing

$$c_i = \frac{\lambda_i}{\Lambda}, \quad x_i = \frac{o_i}{\lambda_i}, \quad i \in \{1, \dots, n\},$$

we then get

$$\left(\sum_{i=1}^n \frac{\lambda_i}{\Lambda} \frac{o_i}{\lambda_i}\right)^2 \leq \sum_{i=1}^n \frac{\lambda_i}{\Lambda} \left(\frac{o_i}{\lambda_i}\right)^2.$$

By several simplifications, we obtain the inequality

$$\frac{1}{\Lambda^2} \left(\sum_{i=1}^n o_i\right)^2 \leq \frac{1}{\Lambda} \sum_{i=1}^n \frac{o_i^2}{\lambda_i}$$

and then (notice that  $\sum_{i=1}^n o_i = l$ )

$$\frac{l^2}{\Lambda} \leq \sum_{i=1}^n \frac{o_i^2}{\lambda_i},$$

with equality again occurring for

$$(1) \quad x_1 = \dots = x_n, \quad \text{tj.} \quad \frac{o_1}{\lambda_1} = \dots = \frac{o_n}{\lambda_n}.$$

This implies that  $S$  the smallest, if and only if (1) holds. This smallest value of  $S$  is  $l^2/(4\pi\Lambda)$ . Now we only need to determine the lengths of the cut parts  $o_i$ . If (1) holds, then clearly  $o_i = k\lambda_i$  for all  $i \in \{1, \dots, n\}$  and certain constant  $k > 0$ . From

$$\sum_{i=1}^n o_i = l \quad \text{and simultaneously} \quad \sum_{i=1}^n o_i = k \sum_{i=1}^n \lambda_i = k\Lambda,$$

we can immediately see that  $k = l/\Lambda$ , i.e.

$$o_i = \frac{\lambda_i}{\Lambda} l, \quad i \in \{1, \dots, n\}.$$

Let's take a look at a specific situation where we are to cut a string of length 1 m into two smaller ones and then create a square and a circle from them so that the sum of their areas is the smallest possible. For a square and a circle (in order), we have (see the example called Isoperimetric quotient)

$$\lambda_1 = \frac{4}{\pi}, \quad \lambda_2 = 1, \quad \text{tj.} \quad \Lambda = \lambda_1 + \lambda_2 = \frac{4+\pi}{\pi}.$$

Then the lengths of the respective parts are (in metres)

$$o_1 = \frac{\frac{4}{\pi}}{\frac{4+\pi}{\pi}} \cdot 1 = \frac{4}{4+\pi} \doteq 0,56, \quad o_2 = \frac{1}{\frac{4+\pi}{\pi}} \cdot 1 = \frac{\pi}{4+\pi} \doteq 0,44.$$

The area of a square with perimeter 0,56 m (with a side of length  $a = 0,14$  m) is 0,0196 m<sup>2</sup> and the area of a circle with perimeter 0,44 m (and radius  $r \doteq 0,07$  m) is approximately 0,0154 m<sup>2</sup>. We can verify that (in m<sup>2</sup>)

$$\frac{l^2}{4\pi\Lambda} = \frac{1}{4(4+\pi)} \doteq 0,035 = 0,0196 + 0,0154.$$

□

A) MATERIAL ON INTEGRATION

**6.D.22.** Based on Sage, evaluate the integrals  $\int f(x) dx$ , where:

- (1)  $f(x) = x^4 + e^x + 5 \ln(x)$ , with  $x > 0$ ;
- (2)  $f(x) = \sqrt{x}(1 + x^3)$ , with  $x > 0$ ;
- (3)  $f(x) = x/\sqrt{x+1}$ , with  $x > -1$ .

○

**6.D.23.** Using any basic formula, and your pencil, determine a primitive function for the following functions:

- (a)  $f(x) = \sqrt{x} \sqrt{x} \sqrt{x}$ , with  $x \in (0, +\infty)$ ;
- (b)  $g(x) = (2^x + 3^x)^2$ , with  $x \in \mathbb{R}$ ;
- (c)  $h(x) = \frac{1}{\sqrt{4-4x^2}}$ , with  $x \in (-1, 1)$ ;
- (d)  $k(x) = \frac{\cos x}{1 + \sin x}$ , with  $x \in (-\frac{\pi}{2}, \frac{3\pi}{2})$ .

Then, confirm your computations via Sage.

○

**6.D.24.** Find a primitive function for the function  $f(x) = e^x + \frac{3}{\sqrt{4-x^2}}$  on the open interval  $(-2, 2)$ .

○

**6.D.25.** Evaluate the integral  $\int \frac{dx}{x(\ln(x))^2 + 2025x}$  with an appropriate substitution.

○

**6.D.26.** Evaluate the integral  $\int \frac{e^x}{e^x + 2024} dx$  with an appropriate substitution.

○

**6.D.27.** Apply integration by parts to compute the integral  $A = \int \cos(3x + 2025)(x + 2026) dx$ , with  $x \in \mathbb{R}$ .

**Solution.** In terms of 6.2.4, it is convenient to set  $F(x) = (x + 2026)$ , such that  $F'(x) = 1$ . Then  $G'(x) = \cos(3x + 2025)$  and  $G$  must have the form  $G(x) = \frac{1}{3} \sin(3x + 2025)$ , such that  $G'(x) = \frac{(3x)' \cos(3x+2025)}{3} = \cos(3x + 2025)$ . Then we get

$$\begin{aligned} A &= \int (x + 2026) \cos(3x + 2025) dx = \int (x + 2026) \left( \frac{\sin(3x + 2025)}{3} \right)' dx = \\ &= \frac{1}{3} (x + 2026) \sin(3x + 2025) - \frac{1}{3} \int (x + 2026)' \sin(3x + 2025) dx = \\ &= \frac{1}{3} (x + 2026) \sin(3x + 2025) - \frac{1}{3} \int \sin(3x + 2025) dx = \\ &= \frac{1}{3} (x + 2026) \sin(3x + 2025) + \frac{1}{3} \cdot \frac{\cos(3x + 2025)}{3} + C, \end{aligned}$$

for some constant  $C$ . A direct verification in Sage occurs as usual:

```
show(integral(cos(3*x+2025)*(x+2026), x).factor())
```

□

**6.D.28.** Compute the integral  $A = \int \sin(2024x) \cos(x) dx$ , formally first and next in Sage.

**Solution.** Recall the identities

$$\sin(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \sin(\beta) \cos(\alpha), \quad \sin(\alpha - \beta) = \sin(\alpha) \cos(\beta) - \sin(\beta) \cos(\alpha).$$

Adding these two relations we get  $2 \sin(\alpha) \cos(\beta) = \sin(\alpha + \beta) + \sin(\alpha - \beta)$ , an identity that one can apply to compute the integral at hand. In particular,

$$\begin{aligned} A &= \int \sin(2024x) \cos(x) dx = \frac{1}{2} \int (\sin(2024x+x) + \sin(2024x-x)) dx = \frac{1}{2} \int \sin(2025x) dx + \frac{1}{2} \int \sin(2023x) dx = \\ &= -\frac{1}{4050} \cos(2025x) - \frac{1}{4046} \cos(2023x) + C. \end{aligned}$$

In Sage to confirm this just type the command `show(integrate(sin(2024*x)*cos(x), x))`. □

**6.D.29.** Compute the integral  $I = \int \sqrt{1-x^2} dx$ , with  $x \in (-1, 1)$ , in two different ways.

**Solution.** In terms of integration by parts (6.2.4), we have  $F(x) = \sqrt{1-x^2}$ ,  $F'(x) = \frac{-x}{\sqrt{1-x^2}}$ ,  $G'(x) = 1$  and  $G(x) = x$ .

Thus, up to a constant we have that

$$\begin{aligned} I &= \int \sqrt{1-x^2} dx = x \sqrt{1-x^2} + \int \frac{x^2}{\sqrt{1-x^2}} dx = x \sqrt{1-x^2} - \int \frac{1-x^2-1}{\sqrt{1-x^2}} dx = \\ &= x \sqrt{1-x^2} - \int \sqrt{1-x^2} dx + \int \frac{1}{\sqrt{1-x^2}} dx = x \sqrt{1-x^2} - \int \sqrt{1-x^2} dx + \arcsin(x). \end{aligned}$$

This implies that

$$2 \int \sqrt{1-x^2} dx = x \sqrt{1-x^2} + \arcsin(x) + C \implies I = \int \sqrt{1-x^2} dx = \frac{1}{2} (x \sqrt{1-x^2} + \arcsin(x)) + C,$$

for some constant  $C \in \mathbb{R}$ .

Let us now compute  $I$  by an appropriate substitution. Set  $x = \sin(t)$  such that  $dx = \cos(t) dt$  and  $x = \arcsin(t)$ . Notice that  $t \in (-\pi/2, \pi/2)$  for  $x \in (-1, 1)$ , and among other things, one has

$$0 < \cos(t) = |\cos(t)| = \sqrt{\cos^2(t)} = \sqrt{1 - \sin^2(t)}.$$

Therefore our integral  $I$  can be written as

$$I = \int \sqrt{1-x^2} dx = \int \sqrt{1-\sin^2(t)} \cos(t) dt = \int \cos^2(t) dt = \frac{1}{2} (t + \sin(t) \cos(t)) + C,$$

for some constant  $C$ , where for the final equality one is based on our previous computation from 6.B.8. Thus

$$I = \frac{1}{2} \left( \sin(t) \sqrt{1-\sin^2(t)} + t \right) + C = \frac{1}{2} \left( x \sqrt{1-x^2} + \arcsin(x) \right) + C, \quad C \in \mathbb{R}.$$

□



**6.D.30.** Compute the limit  $\lim_{x \rightarrow +\infty} \frac{F'(x)}{x^4 + 1}$ , where  $F(x) = \int_1^x (t \sin(t) + 4 \cos(4t)) dt$ .

**Solution.** Let us express  $F$  as  $F(x) = \int_1^x f(t) dt$ , with  $f(t) = t \sin(t) + 4 \cos(4t)$ , for  $t \in \mathbb{R}$ . Since  $f$  is continuous the function  $F$  is differentiable, with  $F'(x) = x \sin(x) + 4 \cos(4x)$  for all  $x \in \mathbb{R}$ . Now for all  $x > 0$  the function  $g(x) = \frac{F'(x)}{x^4 + 1} = \frac{x \sin(x) + 4 \cos(4x)}{x^4 + 1}$  satisfies

$$|g(x)| = \left| \frac{x \sin(x) + 4 \cos(4x)}{x^4 + 1} \right| \leq \frac{|x \sin(x)| + 4 |\cos(4x)|}{x^4 + 1} \leq \frac{|x| + 4}{x^4 + 1} = \frac{x + 4}{x^4 + 1}.$$

It follows that  $\lim_{x \rightarrow +\infty} g(x) = 0$  (since  $\lim_{x \rightarrow +\infty} \frac{x+4}{x^4+1} = 0$ ).  $\square$

**6.D.31.** Suppose that the function  $f : \mathbb{R} \rightarrow \mathbb{R}$  has continuous second derivative on  $\mathbb{R}$  and attains a local extreme at  $x_0 = 1$ . Moreover, assume that the graph  $C_f$  of  $f$  passes through the point  $P = [0, 3] \in \mathbb{R}^2$ . If

$$\int_0^1 (x f''(x) + 4 f'(x)) dx = \alpha$$

for some real number  $\alpha$ , find  $\alpha$  such that  $f(1) = 1/3$ .

**Solution.** By assumption at  $x_0 = 1$  the function  $f$  attains a local extreme, thus  $f'(1) = 0$ . Moreover, we have  $f(0) = 3$ . Thus,

$$\begin{aligned} \int_0^1 (x f''(x) + 4 f'(x)) dx &= \int_0^1 x f''(x) dx + 4 \int_0^1 f'(x) dx = [x f'(x)]_0^1 - \int_0^1 (x') f'(x) dx + 4 [f(x)]_0^1 = \\ &= (1 \cdot f'(1) - 0 \cdot f'(0)) + 3 [f(x)]_0^1 = 0 + 3(f(1) - f(0)) = 3(f(1) - 3) = 3f(1) - 9. \end{aligned}$$

Thus  $3f(1) - 9 = \alpha$ , from where it follows that  $f(1) = (\alpha + 9)/3$ , and since it should be  $f(1) = 1/3$  we get  $\alpha = -8$ .  $\square$

Suppose that for some rational function  $f$  we want to compute the integral  $\int f(\sin(x), \cos(x)) dx$ . Then, usually we apply the substitution method, and here are some hints:

- If  $f(\sin(x), -\cos(x)) = -f(\sin(x), \cos(x))$ , then apply the substitution  $t = \sin(x)$ ;
- If  $f(-\sin(x), \cos(x)) = -f(\sin(x), \cos(x))$ , then set  $t = \cos(x)$ ;
- If  $f(-\sin(x), -\cos(x)) = f(\sin(x), \cos(x))$ , then set  $t = \tan(x)$ ;
- If none of these equalities hold, then try to use the substitution  $t = \tan(x/2)$ .

**6.D.32.** Integrate the following functions, with  $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$  for all the cases:

(a)  $f(x) = \frac{\sin^3(x)}{1 + 4 \cos^2(x) + 3 \sin^2(x)}$ ;

(b)  $g(x) = \frac{1}{1 + \sin^2(x)}$ ;

(c)  $h(x) = \frac{1}{2 - \cos(x)}$ .

**Solution.** For  $f(x)$ , in the denominator it appears the function  $\beta(x) = 1 + 4 \cos^2(x) + 3 \sin^2(x)$ , which can be rewritten as  $\beta(x) = 4 + \cos^2(x)$ , and in the numerator only the sine function to an odd power. Thus, the substitution  $t = \cos(x)$  with  $dt = -\sin(x) dx$ , allows the replacement of all the sines and cosines. Indeed,

$$\begin{aligned} \int \frac{\sin^3(x)}{\beta(x)} dx &= \int \frac{\sin(x) (1 - \cos^2(x))}{4 + \cos^2(x)} dx = \int \frac{-(1 - t^2)}{4 + t^2} dt = \int \left(1 - \frac{5}{4 + t^2}\right) dt = t - \frac{5}{2} \arctan\left(\frac{t}{2}\right) + C \\ &= \cos(x) - \frac{5}{2} \arctan\left(\frac{\cos(x)}{2}\right) + C. \end{aligned}$$

For  $g(x)$ , because both the sine and cosine appear to an even power, we may use the substitution  $t = \tan(x)$ , and hence  $x = \arctan(t)$ . This leads to the relations

$$\sin^2(x) = \frac{\tan^2(x)}{1 + \tan^2(x)} = \frac{t^2}{1 + t^2}, \quad \cos^2(x) = \frac{1}{1 + \tan^2(x)} = \frac{1}{1 + t^2}.$$

Combining them with the relation  $dx = \frac{1}{1+t^2} dt$ , we get

$$\int \frac{dx}{1 + \sin^2(x)} = \int \frac{\frac{1}{1+t^2}}{1 + \frac{t^2}{1+t^2}} dt = \int \frac{1}{1+2t^2} dt = \frac{\sqrt{2}}{2} \arctan(\sqrt{2}t) + C = \frac{\sqrt{2}}{2} \arctan(\sqrt{2} \tan(x)) + C.$$

Finally, for  $h(x)$  one can apply the universal substitution  $t = \tan\left(\frac{x}{2}\right)$ , with

$$\sin(x) = \frac{2t}{1+t^2}, \quad \cos(x) = \frac{1-t^2}{1+t^2}, \quad dx = \frac{2}{1+t^2} dt.$$

In such terms we have

$$\int \frac{dx}{2 - \cos x} = \int \frac{\frac{2}{1+t^2}}{2 - \frac{1-t^2}{1+t^2}} dt = 2 \int \frac{dt}{1+3t^2} = \frac{2\sqrt{3}}{3} \arctan(\sqrt{3}t) + C = \frac{2\sqrt{3}}{3} \arctan\left(\sqrt{3} \tan\left(\frac{x}{2}\right)\right) + C.$$

□

**6.D.33.** Evaluate the integral  $I = \int \frac{1}{\sin(x)} dx$  with  $x \in (0, \frac{\pi}{2})$ , in at least two alternative ways.

**Solution.** We may rewrite the integrand as  $\frac{1}{\sin(x)} = \frac{\sin(x)}{\sin^2(x)} = \frac{\sin(x)}{1 - \cos^2(x)}$ . Thus,  $I = \int \frac{\sin(x)}{1 - \cos^2(x)} dx$  and we may set  $t = \cos(x)$  with  $dt = -\sin(x) dx$ , which gives  $I = -\int \frac{dt}{1-t^2}$ . However, it is easy to prove that  $-\frac{1}{1-t^2} = -\frac{1}{2(t+1)} + \frac{1}{2(t-1)}$ , hence we deduce that

$$\begin{aligned} I &= \frac{1}{2} \int \left( \frac{1}{t-1} - \frac{1}{t+1} \right) dt = \frac{1}{2} \ln|t-1| - \frac{1}{2} \ln|t+1| + C = \frac{1}{2} \ln \left| \frac{t-1}{t+1} \right| + C = \frac{1}{2} \ln \left| \frac{\cos(x)-1}{\cos(x)+1} \right| + C = \\ &= \frac{1}{2} \ln \left| \tan^2\left(\frac{x}{2}\right) \right| + C = \frac{1}{2} \ln \left| \tan\left(\frac{x}{2}\right) \right|^2 + C = \ln \left| \tan\left(\frac{x}{2}\right) \right| + C, \quad C \in \mathbb{R}. \end{aligned}$$

Here we applied the formulas  $\tan\left(\frac{x}{2}\right) = \sqrt{\frac{1-\cos(x)}{1+\cos(x)}}$  and  $|x|^2 = |x^2|$ .

Probably, the quickest way to compute the integral at hand is the substitution  $t = \tan\left(\frac{x}{2}\right)$  (see also 6.D.32), with

$$\sin(x) = \frac{2 \tan\left(\frac{x}{2}\right)}{1 + \tan^2\left(\frac{x}{2}\right)} = \frac{2t}{1+t^2}, \quad dx = \frac{2}{1+t^2} dt.$$

In such terms we immediately get  $I = \int \frac{1}{\frac{2t}{1+t^2}} \cdot \frac{2}{1+t^2} dt = \int \frac{1}{t} dt = \ln|t| + C = \ln \left| \tan\left(\frac{x}{2}\right) \right| + C, C \in \mathbb{R}$ . Notice may exist many other alternatives based on different trigonometric identities. For instance, one may use the identity

$$\tan\left(\frac{x}{2}\right) + \frac{1}{\tan\left(\frac{x}{2}\right)} = \tan\left(\frac{x}{2}\right) + \cot\left(\frac{x}{2}\right) = \frac{2}{\sin(x)}, \quad x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

which you may verify either by your pencil, or by Sage and the command

```
bool(2/sin(x)==(tan(x/2)+(1/tan(x/2)))) for x in RealSet(-pi/2, pi/2)
```

□

*Then next few tasks involve the integration of rational functions.*

**6.D.34.** Consider the rational function  $Q(x) = \frac{1}{x^3 - 1}$ , with  $x \neq 1$ .

(a) Present the decomposition of  $Q$  into partial fractions, and then verify your answer by Sage.

(b) Evaluate the integral  $A = \int Q(x) dx$ .

○

**6.D.35.** Determine the integral  $\int \frac{3x+5}{x^2+4x+8} dx$ , where  $x \in \mathbb{R}$ .

○

**6.D.36.** Compute the indefinite integral of the function  $f(x) = \frac{1}{(x^2+x+1)^2}$ , with  $x \in \mathbb{R}$ .

○

**6.D.37.** Determine the indefinite integral  $\int \frac{dx}{x^3+1}$ , where  $x \in \mathbb{R} \setminus \{-1\}$ .

○

Next, we present additional material related to the integration of irrational expressions. In particular, for a few, we will focus on integrals of the form



$$\int f(x, \sqrt{ax^2 + bx + c}) dx,$$

where  $f$  is a rational expression and we expect  $a \neq 0$ ,  $b^2 - 4ac \neq 0$  for otherwise arbitrary numbers  $a, b, c \in \mathbb{R}$ . In this case one can distinguish two cases, related with the existence/non-existence of real roots for the quadratic polynomial  $ax^2 + bx + c$ .

- If  $a > 0$  and the polynomial  $ax^2 + bx + c$  has real roots  $x_1, x_2$ , use the representation

$$\sqrt{ax^2 + bx + c} = \sqrt{a} \sqrt{(x - x_1)^2 \frac{x - x_2}{x - x_1}} = \sqrt{a} |x - x_1| \sqrt{\frac{x - x_2}{x - x_1}}$$

and set  $t^2 = \frac{x - x_2}{x - x_1}$ .

- If  $a < 0$  and the polynomial  $ax^2 + bx + c$  has real roots  $x_1 < x_2$ , use the representation

$$\sqrt{ax^2 + bx + c} = \sqrt{-a} \sqrt{(x - x_1)^2 \frac{x_2 - x}{x - x_1}} = \sqrt{-a} (x - x_1) \sqrt{\frac{x_2 - x}{x - x_1}}$$

and set  $t^2 = \frac{x_2 - x}{x - x_1}$ .

- Finally, if the polynomial  $ax^2 + bx + c$  does not have real roots (necessarily for  $a > 0$ ), choose the substitution

$$\sqrt{ax^2 + bx + c} = \pm \sqrt{a} \cdot x \pm t$$

with any choice of the signs. Of course, here we choose the signs so that we result to as easy expression to integrate, as possible. As one can expect, for all these cases the corresponding substitutions lead to rational functions.

**6.D.38.** Determine the antiderivative of the following functions:

(a)  $f(x) = \frac{1}{(x + 4)\sqrt{x^2 + 3x - 4}}$ , with  $x \in (-\infty, -4) \cup (1, +\infty)$ ;

(b)  $g(x) = \frac{1}{(x - 1)\sqrt{x^2 + x + 1}}$ , with  $x \neq 1$ .

Next confirm the computations using Sage.

**Solution.** (a) According to the discussion above one can proceed as follows:

$$\int \frac{dx}{(x + 4)\sqrt{x^2 + 3x - 4}} = \int \frac{dx}{(x + 4)\sqrt{(x - 1)(x + 4)}} = \int \frac{dx}{(x + 4)|x + 4| \sqrt{\frac{x - 1}{x + 4}}} = \left| \begin{array}{l} t^2 = \frac{x - 1}{x + 4} \\ x = \frac{5}{1 - t^2} - 4 \\ dx = \frac{10t}{(1 - t^2)^2} dt \end{array} \right| = \int \frac{\frac{10t}{(1 - t^2)^2}}{\left(\frac{5}{1 - t^2}\right) \left|\frac{5}{1 - t^2}\right| t} dt = \int \frac{2}{5} \frac{|1 - t^2|}{1 - t^2} dt = \frac{2}{5} \text{sign}(1 - t^2) \int 1 dt = \frac{2}{5} \text{sign}\left(\frac{5}{x + 4}\right) t + C = \frac{2}{5} \text{sign}(x) \sqrt{\frac{x - 1}{x + 4}} + C.$$

In Sage giving the cell

```
f(x)=1/((x+4)*sqrt(x^2+3*x-4)); show(integral(f(x), x).factor())
```

we get the answer  $\frac{2\sqrt{x^2 + 3x - 4}}{5(x + 4)}$  (recall that  $x^2 + 3x - 4 = (x + 4)(x - 1)$ ). Hence Sage's answer does not contain the sign of  $x$ , but obviously our solution is correct.

(b) Here we see that

$$\int \frac{dx}{(x - 1)\sqrt{x^2 + x + 1}} = \left| \begin{array}{l} \sqrt{x^2 + x + 1} = x + t \\ x^2 + x + 1 = x^2 + 2xt + t^2 \\ x = -\frac{t^2 + 2t - 2}{2t - 1} + 1 \\ dx = \frac{-2(t^2 - t + 1)}{(2t - 1)^2} dt \end{array} \right| = \int \frac{\frac{-2(t^2 - t + 1)}{(2t - 1)^2}}{\frac{-t^2 + 2t - 2}{2t - 1} \frac{t^2 - t + 1}{2t - 1}} dt = \int \frac{2}{t^2 + 2t - 2} dt = \int \left( \frac{\sqrt{3}}{3} \frac{1}{t + 1 - \sqrt{3}} - \frac{\sqrt{3}}{3} \frac{1}{t + 1 + \sqrt{3}} \right) dt = \frac{\sqrt{3}}{3} \ln |t + 1 - \sqrt{3}| - \frac{\sqrt{3}}{3} \ln |t + 1 + \sqrt{3}| + C = \frac{\sqrt{3}}{3} \ln \left| \frac{t + 1 - \sqrt{3}}{t + 1 + \sqrt{3}} \right| + C = \frac{\sqrt{3}}{3} \ln \left| \frac{\sqrt{x^2 + x + 1} - x + 1 - \sqrt{3}}{\sqrt{x^2 + x + 1} - x + 1 + \sqrt{3}} \right| + C.$$

□

**6.D.39.** Using a suitable substitution, compute

$$\int \frac{dx}{x + \sqrt{x^2 + x - 1}} dx, \quad x \in \left(-\infty, \frac{-\sqrt{5}-1}{2}\right) \cup \left(\frac{\sqrt{5}-1}{2}, +\infty\right).$$

**Solution.** Even though the quadratic polynomial under the root has real roots  $x_1, x_2$ , we won't solve this problem by substitution  $t^2 = \frac{x-x_2}{x-x_1}$ . We could proceed that way, but we will rather use a method we introduced for the complex roots case. This is because this method yields a very simple integral of a rational function, as can be seen from the calculation

$$\begin{aligned} \int \frac{dx}{x + \sqrt{x^2 + x - 1}} &= \left| \begin{array}{l} \sqrt{x^2 + x - 1} = x + t \\ x^2 + x - 1 = x^2 + 2xt + t^2 \\ x = \frac{t^2+1}{1-2t} \\ dx = \frac{-2t^2+2t+2}{(1-2t)^2} dt \end{array} \right| = \int \frac{-2t^2 + 2t + 2}{(t+2)(1-2t)} dt = \\ &= \int \left(1 - \frac{2}{t+2} - \frac{1}{2} \frac{1}{t-\frac{1}{2}}\right) dt = t - 2 \ln |t+2| - \frac{1}{2} \ln \left|t - \frac{1}{2}\right| + C = \\ &= \sqrt{x^2 + x - 1} - x - 2 \ln \left(\sqrt{x^2 + x - 1} - x + 2\right) - \frac{1}{2} \ln \left|\sqrt{x^2 + x - 1} - x - \frac{1}{2}\right| + C. \end{aligned}$$

for some constant  $C$ . Notice an undeniable advantage of the recommended substitutions is their universality though: by using them, one can compute all integrals of the respective type.  $\square$

**6.D.40.** For  $x > 0$  determine

$$(a) \int \frac{(2+5x)^3}{\sqrt[4]{x^3}} dx, \quad (b) \int \frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}} dx, \quad (c) \int \frac{1}{\sqrt[4]{1+x^4}} dx.$$

**Solution.** All three given integrals are binomial, i.e. they can be written as

$$\int x^m (a + bx^n)^p dx, \quad \text{for some } a, b \in \mathbb{R}, \text{ and } m, n, p \in \mathbb{Q}.$$

The binomial integrals are usually solved by applying the substitution method. If  $p \in \mathbb{Z}$  (not necessarily  $p < 0$ ), we choose the substitution  $x = t^s$ , where  $s$  is the common denominator of numbers  $m$  and  $n$ ; if  $\frac{m+1}{n} \in \mathbb{Z}$  and  $p \notin \mathbb{Z}$ , we choose  $a + bx^n = t^s$ , where  $s$  is the denominator of number  $p$ ; and if  $\frac{m+1}{n} + p \in \mathbb{Z}$  ( $p \notin \mathbb{Z}$ ,  $\frac{m+1}{n} \notin \mathbb{Z}$ ), we choose  $a + bx^n = t^s x^n$ , where  $s$  is the denominator of  $p$ . In these three cases, a reduction to an integration of a rational function is guaranteed.

Hence we can easily compute

(a)

$$\begin{aligned} \int \frac{(2+5x)^3}{\sqrt[4]{x^3}} dx &= \int x^{-\frac{3}{4}} (2+5x)^3 dx = \left| \begin{array}{l} p \in \mathbb{Z} \\ x = t^4 \\ dx = 4t^3 dt \end{array} \right| = 4 \int (2+5t^4)^3 dt = 4 \int (8 + 60t^4 + 150t^8 + 125t^{12}) dt = \\ &= 4 \left(8t + 12t^5 + \frac{50}{3}t^9 + \frac{125}{13}t^{13}\right) + C = 4 \left(8\sqrt[4]{x} + 12\sqrt[4]{x^5} + \frac{50}{3}\sqrt[4]{x^9} + \frac{125}{13}\sqrt[4]{x^{13}}\right) + C. \end{aligned}$$

(b)

$$\begin{aligned} \int \frac{\sqrt[3]{1+\sqrt[4]{x}}}{\sqrt{x}} dx &= \int x^{-\frac{1}{2}} \left(1+x^{\frac{1}{4}}\right)^{\frac{1}{3}} dx = \left| \begin{array}{l} p \notin \mathbb{Z}, \frac{m+1}{n} \in \mathbb{Z} \\ 1+x^{\frac{1}{4}} = t^3 \\ x = (t^3-1)^4 \\ dx = 12t^2(t^3-1)^3 dt \end{array} \right| = 12 \int t^3(t^3-1) dt = 12 \int t^6 - t^3 dt = \\ &= 12 \left(\frac{t^7}{7} - \frac{t^4}{4}\right) + C = 12\sqrt[3]{(1+\sqrt[4]{x})^4} \left(\frac{1+\sqrt[4]{x}}{7} - \frac{1}{4}\right) + C. \end{aligned}$$

(c)

$$\begin{aligned} \int \frac{1}{\sqrt[4]{1+x^4}} dx &= \int (1+x^4)^{-\frac{1}{4}} dx = \left| \begin{array}{l} p \notin \mathbb{Z}, \frac{m+1}{n} \notin \mathbb{Z}, \frac{m+1}{n} + p \in \mathbb{Z} \\ 1+x^4 = t^4 x^4 \\ x = (t^4-1)^{-\frac{1}{4}} \\ dx = -t^3(t^4-1)^{-\frac{5}{4}} dt \end{array} \right| = - \int \frac{t^2}{t^4-1} dt = - \int \frac{t^2}{(t-1)(t+1)(t^2+1)} dt = \\ &= -\frac{1}{4} \int \left(\frac{1}{t-1} - \frac{1}{t+1} + \frac{2}{t^2+1}\right) dt = -\frac{1}{4} (\ln |t-1| - \ln |t+1| + 2 \arctg t) + C = \\ &= -\frac{1}{4} \left[ \ln \frac{\sqrt[4]{\frac{1}{x^4}+1}-1}{\sqrt[4]{\frac{1}{x^4}+1}+1} + 2 \arctg \left(\sqrt[4]{\frac{1}{x^4}+1}\right) \right] + C. \end{aligned}$$

□

Next, we focus on definite integrals and related applications.

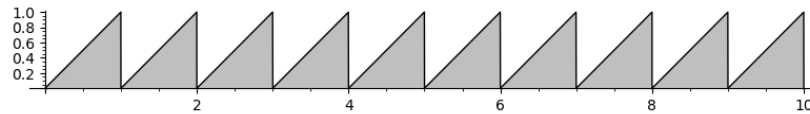
**6.D.41.** Compute the definite integrals given below:

$$A = \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \frac{\sin t}{1 - \cos^2 t} dt, \quad B = \int_0^{\ln 2} \frac{dx}{e^{2x} - 3e^x}, \quad C = \int_0^{\frac{\pi}{2}} \sin(x) \sin(2x) dx.$$

○

**6.D.42.** Use 6.B.36, to compute the integral  $\int_0^{50} \{x\} dx$ , where  $\{ \}$  is the fractional part function.

**Solution.** (c) Recall by 5.E.57 that the fractional part function is defined by  $\{x\} = x - \lfloor x \rfloor$ , and clearly this function is periodic with period  $T = 1$ . The integral at hand corresponds to the grey area in the following figure, which equals to  $10 \cdot \frac{1}{2} = 5$  (obviously, this is the area of 10 half square boxes with sides of length 1).



Indeed,

$$\int_0^{10} \{x\} dx = 10 \int_0^1 \{x\} dx = 10 \int_0^1 x dx = 10 \left[ \frac{x^2}{2} \right]_0^1 = 5.$$

For your convenience, here is the code used in Sage to produce the previous figure and confirm the result:

```
fract(x)=x-floor(x)
p=plot(fract, x, 0, 10, color="black", aspect_ratio=1, fill=True, fillcolor="grey", figsize=8); show(p)
integral(fract, x, 0, 10)
```

□

**6.D.43.** For some parameter  $\alpha \in \mathbb{R}$ , consider the function

$$f(x) = \begin{cases} e^x \cos(2x), & \text{if } x \in [-\pi/2, 0], \\ \sqrt{\alpha + \sin(4x)}, & \text{if } x \in (0, 3\pi/8]. \end{cases}$$

(a) Determine  $\alpha$  such that  $f$  is continuous for any  $x$  in its domain. Then use your answer and Sage to execute the following:

- Indicate the whole domain of  $f$  but also the domains of the two components of  $f$ .
- Sketch the graph of  $f$  via the piecewise method in Sage (see also 5.C.5).

(b) Compute the area in between the graph of  $f$  and the  $x$ -axis, bounded from the vertical lines  $x = \pm \frac{\pi}{4}$ .

(c) Use Sage in order to

- Confirm your result in (b), in a “manual way”.
- Compute the error that one obtains, if any, integrating the function  $f$ , as introduced via the piecewise method.

**Solution.** At  $x_0 = 0$  we see that  $\lim_{x \rightarrow 0^-} f(x) = \lim_{x \rightarrow 0^-} (e^x \cos(2x)) = 1 = f(0)$ . Hence we need

$$\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^+} \sqrt{\alpha + \sin(4x)} = 1 \iff \sqrt{\alpha + \lim_{x \rightarrow 0^+} \sin(4x)} = 1,$$

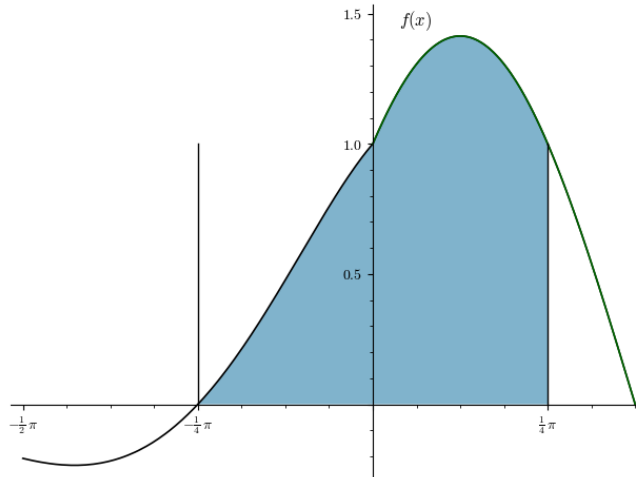
that is,  $\sqrt{\alpha + 0} = 1$  and hence  $\alpha = 1$ . Thus  $f$  is continuous everywhere in its domain if and only if  $\alpha = 1$  (obviously,  $f$  is continuous on the intervals  $[-\pi/2, 0)$  and  $(0, 3\pi/8]$ ). From now fix  $\alpha = 1$  and observe that  $f(-\pi/4) = 0$  and  $f(3\pi/8) = 0$ .

Sage has built-in commands to indicate the domain of a piecewise function and of its components, given by `domain` and `domains`, respectively. Hence we may type

```
f1(x)=(e^x)*cos(2*x); f2(x)=sqrt(1+sin(4*x))
f=piecewise([[(-pi/2, 0), f1(x)], [(0, 6*pi/16), f2(x)]])
f.domain(); f.domains()
```

Sage returns the obvious answers, i.e.,  $(-1/2 * \pi, 0) \cup (0, 3/8 * \pi)$ , and  $((-1/2 * \pi, 0), (0, 3/8 * \pi))$ , respectively.

Let us now present the graph of  $f$ , where we have included the two vertical lines from the question in (b), and shaded the region whose area needs to be computed.



The code to construct this figure relies on many different steps, and some advanced options from 2D-graphics in Sage, as for example the command `polygon` to shade the region in question.

In general, when one wants to integrate a function  $f$  from  $a$  to  $b$  and compute the area between  $f$  and the  $x$ -axis, with bounds the vertical lines passing from  $a, b$ , then can follow the summary given here:

- Step 1 introduce the function  $f$ ;  
 Step 2 declare the points  $a, b$   
 Step 3 sketch the graph of  $f$ , e.g., add the command

```
q= plot(f, (x, a-0.5, b+0.5), thickness=2)
```

- Step 4 graph the two vertical lines, which can be done by adding the syntax

```
q+= line([(a, 0), (a, f(a))], color='black')
q+= line([(b, 0), (b, f(b))], color='black')
```

- Step 5 shade the area at hand, via the `polygon` method:

```
q += polygon([(a, 0), (a, f(a))] + [(x, f(x)) for x in [a, a+0.005, ..., b]] + [(b, 0), (a, 0)],
            rgbcolor=(0.2, 0.4, 0.7), aspect_ratio="automatic")
```

- Step 6 produce the figure, which can be done by adding `show(q)` (we may specify further options here, as `ymin`, `ymax`, `ticks`, `figsize`, and other, hence instead one may type `q.show()`, and add inside the parentheses the appropriate options).

For our example this method becomes a few more complicated, since we have a piecewise function, and hence we use first the `piecewise` method to introduce  $f$ . Here is the full implementation combining both methods.

```
f1(x)=(e^x)*cos(2*x); f2(x)=sqrt(1+sin(4*x))
f=piecewise([( (-pi/2, 0), f1(x)), [(0, 6*pi/16), f2(x) ]])
p=plot(f(x), x, -pi/2, 6*pi/16, color="black", thickness=1.2)
p+=plot(f2(x), x, 0, 6*pi/16, color="darkgreen", thickness=1.2)
p+= text (" $f(x)$ ", (pi/16, 1.47), color="black", fontsize="12")
a=-pi/4; b=pi/4
p+=line([(a, 0), (a, 1)], color="black")
p+=line([(b, 0), (b, f(N(b)))]), color="black")
p+=polygon([( (-pi/4, 0), (-pi/4, f(N(-pi/4))) )]+[(x, f(N(x)))]
            for x in [-pi/4, -pi/4+0.005, ..., pi/4]]+[(b, 0), (a, 0)],
            rgbcolor=(0.5, 0.7, 0.8), aspect_ratio="automatic")
p.show(ticks=pi/4, tick_formatter=pi, ymin=-0.25, ymax=1.5)
```

Notice this technique requires the values introduced in  $f$  to be in numerical expression, and this is the reason behind the appearance of  $f(N(b))$ ,  $f(N(x))$ , etc. Otherwise Sage returns errors. However, as we will see below this method produces a small error in the computation of the given area; this is because the `piecewise` method uses approximations of  $a, b$ , which effect on the integration.

(b) Let us denote by  $E$  the shaded area. By definition we have

$$E = \int_{-\pi/4}^{\pi/4} |f(x)| dx = \int_{-\pi/4}^{\pi/4} f(x) dx = \int_{-\pi/4}^0 (e^x \cos(2x)) dx + \int_0^{\pi/4} \sqrt{1 + \sin(4x)} dx = E_1 + E_2.$$

For the first integral, recall that  $\left(\frac{\sin(2x)}{2}\right)' = \cos(2x)$  and  $\left(-\frac{\cos(2x)}{2}\right)' = \sin(2x)$ . Thus, by applying integration by parts twice we get

$$\begin{aligned} E_1 &= \int_{-\pi/4}^0 e^x \left(\frac{\sin(2x)}{2}\right)' dx = \frac{1}{2} [e^x \sin(2x)]_{-\pi/4}^0 - \frac{1}{2} \int_{-\pi/4}^0 e^x \sin(2x) dx = \\ &= \frac{1}{2} \left( e^0 \cdot \sin(0) - e^{-\pi/4} \cdot \sin\left(-\frac{2\pi}{4}\right) \right) + \frac{1}{4} \int_{-\pi/4}^0 e^x (\cos(2x))' dx = \\ &= \frac{1}{2} e^{-\pi/4} \cdot \sin\left(\frac{\pi}{2}\right) + \frac{1}{4} [e^x \cos(2x)]_{-\pi/4}^0 - \frac{1}{4} \int_{-\pi/4}^0 e^x \cos(2x) dx. \end{aligned}$$

Thus

$$E_1 = \frac{1}{2} e^{-\pi/4} + \frac{1}{4} \left( e^0 \cdot \cos(0) - e^{-\pi/4} \cdot \cos\left(-\frac{2\pi}{4}\right) \right) - \frac{1}{4} E_1,$$

or equivalently  $E_1 = \frac{2e^{-\pi/4} + 1}{5} \approx 0.382375$ .

For  $E_2$ , we use the identities  $\sin(2y) = 2 \sin(y) \cos(y)$  and  $\sin^2(y) + \cos^2(y) = 1$ , for any  $y \in \mathbb{R}$ , where we replace  $y$  by  $2x$ . Thus we can write

$$\begin{aligned} E_2 &= \int_0^{\pi/4} \sqrt{\sin^2(2x) + \cos^2(2x) + 2 \sin(2x) \cos(2x)} dx = \int_0^{\pi/4} \sqrt{(\sin(2x) + \cos(2x))^2} dx = \\ &= \int_0^{\pi/4} (\sin(2x) + \cos(2x)) dx = \left[ -\frac{\cos(2x)}{2} + \frac{\sin(2x)}{2} \right]_0^{\pi/4} = \frac{1}{2} + \frac{1}{2} = 1. \end{aligned}$$

Thus all together  $E = E_1 + E_2 = \frac{2e^{-\pi/4} + 1}{5} + 1 = \frac{2e^{-\pi/4} + 6}{5} \approx 1.382375$ .

(c) A confirmation of the result above, occurs easily, without using the piecewise method to introduce  $f$ , and from the mathematical point of view it relies on the relation  $E = E_1 + E_2$ . Hence we can type

```
f1(x)=e^(x)*cos(2*x); f2(x)=sqrt(1+sin(4*x))
show(integral(f1(x), x, -pi/4, 0))
show(integral(f2(x), x, 0, pi/4))
```

and this is what we mean “manually”. Execute this cell yourself to check Sage’s output.

On the other hand, introducing the function  $f$  (in Sage) via the piecewise method, gives us the ability to directly integrate  $f$ , as follows:

```
f1(x)=(e^x)*cos(2*x); f2(x)=sqrt(1+sin(4*x))
f=piecewise([[(-pi/2, 0), f1(x)], [(0, 6*pi/16), f2(x)]])
f.integral(x, N(-pi/4), N(pi/4))
```

However, since this still requires to introduce the certain values  $a = -\pi/4$  and  $b = \pi/4$  numerically, some error will appear. Indeed, executing this block, Sage prints out 1.16777341450385, which differs from the result in (b). The same time, we cannot replace in this certain example the last command by `f.integral(x, -pi/4, pi/4)`, which is a pitfall of Sage.  $\square$

**6.D.44.** Compute the area  $S$  of a figure composed of two parts of plane bounded by lines  $x = 0$ ,  $x = 1$ ,  $x = 4$ , the  $x$  axis and the graph of a function

$$y = \frac{1}{\sqrt[3]{x-1}}.$$

**Solution.** First realize that

$$\frac{1}{\sqrt[3]{x-1}} < 0, \quad x \in [0, 1), \quad \frac{1}{\sqrt[3]{x-1}} > 0, \quad x \in (1, 4]$$

and

$$\lim_{x \rightarrow 1^-} \frac{1}{\sqrt[3]{x-1}} = -\infty, \quad \lim_{x \rightarrow 1^+} \frac{1}{\sqrt[3]{x-1}} = +\infty.$$

The first part of the figure (below the  $x$  axis) is thus bounded by the curves

$$y = 0, \quad x = 0, \quad x = 1, \quad y = \frac{1}{\sqrt[3]{x-1}}$$

with an area given by the improper integral

$$S_1 = -\int_0^1 \frac{1}{\sqrt[3]{x-1}} dx;$$

while the second part (above the  $x$  axis), which is bounded by the curves

$$y = 0, \quad x = 1, \quad x = 4, \quad y = \frac{1}{\sqrt[3]{x-1}},$$

has an area of

$$S_2 = \int_1^4 \frac{1}{\sqrt[3]{x-1}} dx.$$

Since

$$\int \frac{1}{\sqrt[3]{x-1}} dx = \frac{3}{2} \sqrt[3]{(x-1)^2} + C,$$

the sum  $S_1 + S_2$  can be gotten as

$$S = - \lim_{x \rightarrow 1^-} \left( \frac{3}{2} \sqrt[3]{(x-1)^2} - \frac{3}{2} \right) + \lim_{x \rightarrow 1^+} \left( \frac{3}{2} \sqrt[3]{9} - \frac{3}{2} \sqrt[3]{(x-1)^2} \right) = \frac{3}{2} (1 + \sqrt[3]{9}).$$

We have shown among other things, that the given figure has a finite area, even though it's unbounded (both from the top and the bottom). (If we approach  $x = 1$  from the right, eventually from the left, its altitude grows beyond measure.) Recall here the indefinite expression of type  $0 \cdot \infty$ . Namely, the figure is bounded if we limit ourselves to  $x \in [0, 1 - \delta] \cup [1 + \delta, 4]$  for an arbitrarily small  $\delta > 0$ .  $\square$

**6.D.45.** Determine the surface and volume of a circular paraboloid created by rotating a part of the parabola  $y = 2x^2$  for  $x \in [0, 1]$  around the  $y$  axis.

**Solution.** The formulas stated in the texts are true for rotating the curves around the  $x$  axis! Hence it's necessary either to integrate the given curve with respect to variable  $y$ , or to transform. This gives

$$V = \int_0^2 \frac{x}{2} dx = \pi \quad \text{and} \quad S = 2\pi \int_0^2 \sqrt{\frac{x}{2}} \sqrt{1 + \frac{1}{8x}} dx = 2\pi \int_0^2 \sqrt{\frac{x}{2} + \frac{1}{16}} dx = \pi \frac{17\sqrt{17} - 1}{24}.$$

$\square$

A) MATERIAL ON SEQUENCES, SERIES AND LIMIT PROCESSES

*We begin with additional material related to uniform convergence of sequence of functions and of series.*

**6.D.46. Uniform convergence.** Let  $(f_n)$  be a sequence of functions defined on an interval  $I$  such that  $\lim_{n \rightarrow \infty} f_n(x) = f(x)$  for all  $x \in I$ . Prove that  $(f_n)$  uniformly converges to  $f$  on  $I$  if and only if  $\lim_{n \rightarrow \infty} a_n = 0$ , where  $a_n := \sup_{x \in I} |f_n(x) - f(x)|$ .

**Solution.** Suppose that  $(f_n)$  is uniformly convergent to  $f$  on  $I$ . Then, given  $\varepsilon > 0$  there exists an integer  $N$  such that for any  $n \geq N$  and  $x \in I$  we have  $|f_n(x) - f(x)| < \varepsilon$ . Therefore we also have  $a_n := \sup_{x \in I} |f_n(x) - f(x)| \leq \varepsilon$  for all  $n \geq N$  and thus  $a_n \rightarrow 0$  as  $n \rightarrow \infty$ . Conversely, assume that  $\lim_{n \rightarrow \infty} a_n = 0$ , where  $a_n$  is defined as above. Then, for any  $\varepsilon > 0$  there exists an integer  $N$  such that  $a_n < \varepsilon$  for all  $n \geq N$ , that is,

$$\sup_{x \in I} |f_n(x) - f(x)| < \varepsilon, \quad \text{for all } n \geq N.$$

By the definition of supremum,

$$|f_n(x) - f(x)| \leq \sup_{x \in [a,b]} |f_n(x) - f(x)| < \varepsilon,$$

for sufficiently large  $n$  and for all  $x \in I$ . Thus  $(f_n)$  uniformly converges to  $f$  on  $I$ .  $\square$

**6.D.47. Natural logarithm.** Expand the natural logarithm  $f(x) = \ln(1+x)$  into a power series around 0 and 1, and next determine all  $x \in \mathbb{R}$  for which these series converge.

**Solution.** First we will determine the expansion around point 0. To expand a function into a power series at a given point is the same as to determine its Taylor expansion at that point. We can easily see that

$$f^{(n)}(x) = [\ln(x+1)]^{(n)} = (-1)^{n+1} \frac{(n-1)!}{(x+1)^n},$$

so after computing the derivatives at zero, we have

$$f(x) = \ln(x+1) = \ln 1 + \sum_{n=1}^{\infty} a_n x^n,$$

where the coefficients  $a_n$  are given by

$$a_n = \frac{(-1)^{n+1} (n-1)!}{n!} = \frac{(-1)^{n+1}}{n}.$$



Therefore, one can write

$$\begin{aligned} f(x) = \ln(x+1) &= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \dots \\ &= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n. \end{aligned}$$

For the radius of convergence, we can then use the limit of the quotient of the following coefficients of terms of the power series:

$$r = \frac{1}{\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|} = \frac{1}{\lim_{n \rightarrow \infty} \frac{\frac{1}{n+1}}{\frac{1}{n}}} = 1.$$

Hence the series converges for arbitrary  $x \in (-1, 1)$ . For  $x = -1$  we get the harmonic series (with a negative sign), while for  $x = 1$  we get the alternating harmonic series, which converges by the Leibniz criterion. Thus the given series converges exactly for  $x \in (-1, 1]$ .

Analogously, for the expansion at point 1, we get

$$\begin{aligned} f(x) = \ln(x+1) &= \ln(2) + \frac{1}{2}(x-1) - \frac{1}{8}(x-1)^2 \\ &\quad + \frac{1}{3 \cdot 2^3}(x-1)^3 - \dots \\ &= \ln(2) + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n \cdot 2^n} (x-1)^n, \end{aligned}$$

and for the radius of convergence of this series we get

$$r = \frac{1}{\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right|} = \frac{1}{\lim_{n \rightarrow \infty} \frac{\frac{1}{2^{n+1}(n+1)}}{\frac{1}{2^n n}}} = 1.$$

□

- 6.D.48.** Expand the function  $\cos^2(x)$  into a power series at 0, and determine for which  $x \in \mathbb{R}$  it converges.
- 6.D.49.** Expand the function  $\sin^2(x)$  into a power series at 0 and determine for which  $x \in \mathbb{R}$  it converges.
- 6.D.50.** Expand the function  $\ln(x^3 + 3x^2 + 3x + 1)$  into a power series at 0 and determine for which  $x \in \mathbb{R}$  it converges.
- 6.D.51.** Expand the function  $\ln(\sqrt{x})$  into a power series at 1 and determine for which  $x \in \mathbb{R}$  it converges.

Now is the time to demonstrate various ways to combine the theory of Taylor series with differentiation and integration. Further similar tasks are presented in Section D (see for example 6.D.58). To make the description more engaging, we will also utilize Sage.



- 6.D.52.** On the interval of convergence  $(-1, 1)$ , determine the sum of the series

$$\sum_{n=1}^{\infty} n(n+1)x^n.$$

Next confirm your answer in Sage.

**Solution.** We have

$$\begin{aligned} \sum_{n=1}^{\infty} n(n+1)x^n &= \sum_{n=1}^{\infty} n(x^{n+1})' = \left( \sum_{n=1}^{\infty} n x^{n+1} \right)' \\ &= \left( \sum_{n=1}^{\infty} n x^{n-1} x^2 \right)' = \left[ x^2 \sum_{n=1}^{\infty} (x^n)' \right]' \\ &= \left[ x^2 \left( \sum_{n=1}^{\infty} x^n \right)' \right]' = \left[ x^2 \left( -1 + \sum_{n=0}^{\infty} x^n \right)' \right]' \\ &= \left[ x^2 \left( -1 + \frac{1}{1-x} \right)' \right]' = \left[ x^2 \cdot \frac{1}{(1-x)^2} \right]' = \frac{2x}{(1-x)^3}, \end{aligned}$$

for all  $x \in (-1, 1)$ . To confirm the result in Sage use the `sum` command, as follows:

```
var("n")
show(sum(n*(n+1)*x^n, n, 1, oo).factor())
```

□

**6.D.53.** Use the Taylor series of  $f(x) = \cos(\sqrt{x})$  centered at 0 to approximate the integral  $I = \int_0^1 \cos(\sqrt{x}) dx$ . Then use Sage to compute the real value of  $I$ , and compare the two results. ○

**6.D.54. Taylor series of arctan.** Consider the function  $f(x) = 1/(1+x^2)$ ,  $x \in (-1, 1)$ . Use the Taylor series of  $f$  centered at 0 and the notion of definite integrals to compute the power series of  $\arctan(x)$  around 0.

**Solution.** Recall that for  $|x| < 1$  we have  $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$ . Thus the replacement of  $x$  by  $-x^2$  gives (see also 5.E.142)

$$\frac{1}{1+x^2} = \sum_{n=0}^{\infty} (-1)^n x^{2n}, \quad |x| < 1.$$

Hence, for all  $t \in (-1, 1)$  we have

$$\arctan'(t) = \frac{1}{1+t^2} = \sum_{n=0}^{\infty} (-t^2)^n = \sum_{n=0}^{\infty} (-1)^n t^{2n}.$$

Now, for all  $x \in (-1, 1)$  we have  $\int_0^x \arctan'(t) dt = \arctan(x) - \arctan(0) = \arctan(x)$ . On the other side

$$\int_0^x \left( \sum_{n=0}^{\infty} (-1)^n t^{2n} \right) dt = \sum_{n=0}^{\infty} \left( (-1)^n \int_0^x t^{2n} dt \right) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1},$$

and thus  $\arctan(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}$  for all  $x \in (-1, 1)$ . □

**6.D.55.** Find the Maclaurin series of the function

$$A(x) = \int_0^x t \cos(t^2) dt.$$

**Solution.** The Maclaurin series of  $\cos(t)$  is

$$1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \dots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} t^{2n}.$$

Using this we see that

$$t \cos(t^2) = t \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{(2n)!} = \sum_{n=0}^{\infty} (-1)^n \frac{t^{4n+1}}{(2n)!}.$$

Therefore,

$$\begin{aligned} A(x) &= \int_0^x t \cos(t^2) dt = \int_0^x \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} t^{4n+1} dt \\ &= \sum_{n=0}^{\infty} \left[ \frac{(-1)^n}{(2n)!} \int_0^x t^{4n+1} dt \right] \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(4n+2)(2n)!} x^{4n+2}. \end{aligned}$$

Let us write down the first few terms of  $A$ ,

$$A(x) = \frac{1}{2}x^2 - \frac{1}{12}x^6 + \frac{1}{240}x^{10} - \frac{1}{10080}x^{14} + \dots$$

A confirmation of this expression via Sage is based on the method presented in 6.A.17. However, one should now combine this method with the command `integral(g(t), t, 0, x)`, to introduce a function of the form  $\int_0^x g(t) dt$ , where  $g(t) = t \cos(t^2)$ .

Thus the implementation goes as follows:

```
var("t"); assume(x>0)
f(x)=integral(t*cos(t^2), t, 0, x)
tf=taylor(f, x, 0, 6)
T=tf.power_series(QQ); show(T)
```

□

**6.D.56.** For the convergent series

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{\sqrt{n+100}}$$

estimate the error of the approximation of its sum by the partial sum  $s_{9999}$ .

○

**6.D.57.** Approximate the expression given below with an error lesser than  $1/10$ :

$$\int_1^2 \left( x - \frac{\cos^{10}(x)}{10} \right) \ln(x) dx.$$

○

**6.D.58.** Find the Maclaurin series of the function  $F(x) = \int_0^x \frac{\sin(t)}{t} dt$ , with  $x \in \mathbb{R}$ . Next check your answer via Sage, and plot the graph of the function  $F$  for certain  $x$  in the domain of  $F$ .

**Solution.** Recall that  $\sin(t) = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \dots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} t^{2n+1}$  for all  $t \in \mathbb{R}$ , hence

$$\frac{\sin(t)}{t} = 1 - \frac{t^2}{3!} + \frac{t^4}{5!} - \dots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} t^{2n}.$$

By integrating this series we get the series expansion of  $F(x)$ , i.e.,

$$\begin{aligned} F(x) &= \int_0^x \frac{\sin(t)}{t} dt = \int_0^x \left[ \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} t^{2n} \right] dt = \sum_{n=0}^{\infty} \left[ \frac{(-1)^n}{(2n+1)!} \int_0^x t^{2n} dt \right] = \sum_{n=0}^{\infty} \left[ \frac{(-1)^n}{(2n+1)!} \frac{t^{2n+1}}{(2n+1)} \right]_0^x \\ &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)(2n+1)!} x^{2n+1}. \end{aligned}$$

Hence the series expansion of  $F$  is given by

$$F(x) = x - \frac{x^3}{3 \cdot 3!} + \frac{x^5}{5 \cdot 5!} - \dots + \frac{(-1)^n}{(2n+1)(2n+1)!} x^{2n+1} + \dots$$

We can easily check this result in Sage by the method described in 6.A.17 and 6.D.55, that is,

```
var("t"); assume(x>0); F(x)=integral(sin(t)/t, t, 0, x)
tF=taylor(F, x, 0, 7); T=tF.power_series(QQ); show(T)
```

This cell prints out the following expression:  $x - \frac{1}{18}x^3 + \frac{1}{600}x^5 - \frac{1}{35280}x^7 + O(x^8)$ , hence verifies our answer. In order to plot the graph of  $F$  for certain  $x$  we can for example add in the above cell the command `show(plot(F, x, -10, 10))`.

**6.D.59.** Find the function  $f$  to which, for  $x \in \mathbb{R}$ , the sequence of functions

$$f_n(x) = \frac{n^2 x^3}{n^2 x^2 + 1}, \quad n \in \mathbb{N}.$$

converges. Is this convergence uniform on  $\mathbb{R}$ ?

**6.D.60.** Does the series

$$\sum_{n=1}^{\infty} \frac{nx}{n^4 + x^2}, \quad \text{kde } x \in \mathbb{R},$$

converge uniformly on the real line?

**6.D.61.** Approximate

(a) the cosine of ten degrees with accuracy of at least  $10^{-5}$ ;

(b) the definite integral  $\int_0^{1/2} \frac{dx}{x^4 + 1}$  with accuracy of at least  $10^{-3}$ .

**6.D.62.** Determine the power series centered at  $x_0 = 0$  of the function

$$f(x) = \int_0^x e^{t^2} dt, \quad x \in \mathbb{R}.$$

**6.D.63.** Using the integral test, find the values  $a > 0$  for which the series

$$\sum_{n=1}^{\infty} \frac{1}{n^a}$$

converges.

**6.D.64.** For which  $x \in \mathbb{R}$  does the series

$$\sum_{n=1}^{\infty} \frac{\ln(n!)}{n^x}$$

converge?

**6.D.65.** Determine whether the series

$$\sum_{n=1}^{\infty} (-1)^{n-1} \tan \frac{1}{n\sqrt{n}}$$

converges absolutely, converges conditionally, diverges to  $+\infty$ , diverges to  $-\infty$ , or none of the above. (such a series is sometimes said to be oscillating).

**6.D.66.** Calculate the series

$$\sum_{n=1}^{\infty} \frac{1}{n \cdot 3^n}$$

with the help of an appropriate power series.

Solutions of the exercises

**6.A.5.** For all  $x \in (-1, 1)$  we have  $\ln \frac{1+x}{1-x} = \ln(1+x) - \ln(1-x)$ . Therefore, it is useful to consider the auxiliary function

$$\varphi(x) := \ln(ax+1), \quad x \in (-1, 1), \quad a = \pm 1.$$

For this function we easily compute

$$\varphi'(x) = \frac{a}{ax+1}, \quad \varphi''(x) = \frac{-a^2}{(ax+1)^2}, \quad \varphi^{(3)}(x) = \frac{2a^3}{(ax+1)^3}, \quad \varphi^{(4)}(x) = \frac{-6a^4}{(ax+1)^4},$$

for all  $x \in (-1, 1)$ . Motivated by these formulas, one may guess that

$$\varphi^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!a^n}{(ax+1)^n}, \quad x \in (-1, 1), \quad n \in \mathbb{N}. \quad (*)$$

Let us use the principle of mathematical induction to prove (\*). As we have seen, it holds for  $n = 1, 2, 3, 4$  and we assume that (\*) holds for some other  $k \in \mathbb{N}$ . Then, a direct computation shows that

$$\varphi^{(k+1)}(x) = (\varphi^{(k)}(x))' = \left( \frac{(-1)^{k-1}(k-1)!a^k}{(ax+1)^k} \right)' = \frac{(-1)^{k-1}(k-1)!a^k(-k)a}{(ax+1)^{k+1}} = \frac{(-1)^k k! a^{k+1}}{(ax+1)^{k+1}},$$

and hence the relation (\*) is true for all  $n \in \mathbb{N}$ . Let us now return back to our initial task. By (\*) we may compute the  $n$ th derivative of  $\ln(1 \pm x)$ , that is,

$$\ln^{(n)}(1+x) = \frac{(-1)^{n-1}(n-1)!}{(x+1)^n}, \quad \ln^{(n)}(1-x) = -\frac{(n-1)!}{(-x+1)^n}, \quad x \in (-1, 1).$$

Therefore, we finally deduce that

$$\left( \ln \frac{1+x}{1-x} \right)^{(n)} = (n-1)! \left( \frac{1}{(1-x)^n} - \frac{(-1)^n}{(1+x)^n} \right)$$

for all  $x \in (-1, 1)$  and  $n \in \mathbb{N}$ .

**6.A.6.** The answer is  $f^{(12)}(x) = 2^{12} e^{2x} + \cos(x)$ .

**6.A.7.** The answer is  $f^{(26)}(x) = -\sin(x) + 2^{26} e^{2x}$ .

**6.A.8.** Recall that the third-order Taylor expansion of a function  $f$  around the point  $a = 0$ , is given by

$$T_0^3 f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f^{(3)}(0)}{6}x^3.$$

For the sine function  $f(x) = \sin(x)$  we have  $f'(0) = \cos(0) = 1$ ,  $f^{(2)}(0) = -\sin(0) = 0$ ,  $f^{(3)}(0) = -\cos(0) = -1$ , and  $f(0) = 0$ . Thus we get

$$T_0^3 \sin(x) = x - \frac{1}{6}x^3 = x - \frac{x^3}{3!}.$$

In a similar way we obtain

$$T_0^3 \cos(x) = 1 - \frac{x^2}{2}, \quad T_0^3 e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6}, \quad T_0^3 \ln(x+1) = x + \frac{x^2}{2} - \frac{x^3}{3}.$$

**6.A.10.** The answer is

$$T_1^4 f(x) = 2(x-1) - (x-1)^2 + \frac{2}{3}(x-1)^3 - \frac{1}{2}(x-1)^4.$$

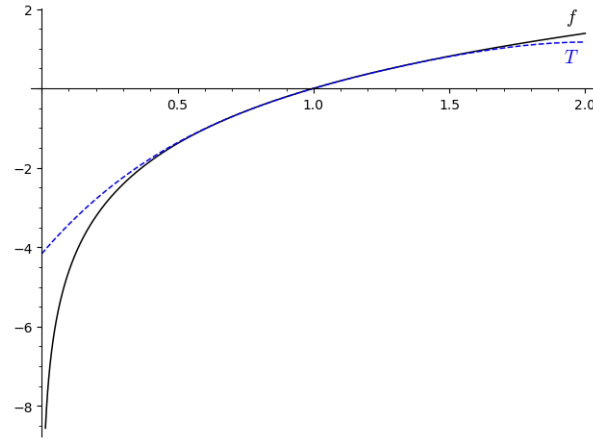
To obtain this expression via Sage, we may type

```
f(x)=ln(x^2); T(x)=taylor(f(x), x, 1, 4); show(T(x))
```

Adding in this cell the command

```
p=plot(f(x), x, 0, 2, color="black", thickness=1.1)
p+=plot(T(x), x, 0, 2, linestyle="--", thickness=1.1);
p+=text(r"$f$", (1.95,1.8), fontsize=14, color="black")
p+=text(r"$T$", (1.95,0.8), fontsize=14, color="blue")
show(p)
```

one can produce the required graphs which we present here (the graph of the Taylor polynomial is coloured blue).



Observe that the approximation is accurate enough. In fact, try to verify yourself that as the degree of the Taylor polynomials increases, the approximations become more accurate.

**6.A.13.** Consider the exponential function  $f(x) = e^x$  restricted on the closed interval  $[0, x]$ . The third-order degree Taylor polynomial of  $f(x) = e^x$ , centered at  $a = 0$  has the form

$$T_0^3 e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}.$$

Now,  $f$  is a smooth function and according to the theorem in 6.1.3 there exists some  $c \in (0, x)$  such that

$$f(x) = e^x = T_0^3 e^x + R(x), \quad R(x) = \frac{x^{3+1}}{(3+1)!} f^{(3+1)}(c).$$

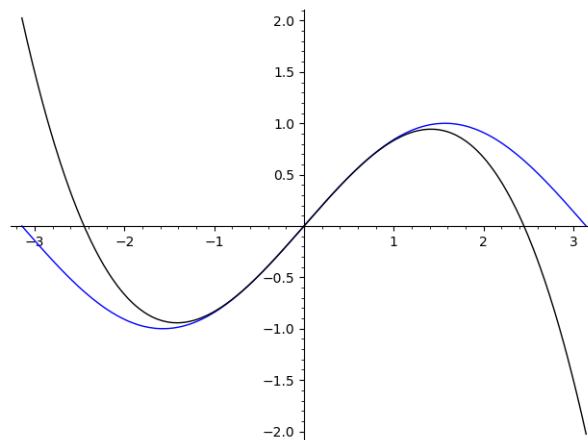
We have  $f^{(4)}(x) = e^x$  and hence we see that  $R(x) = \frac{x^4}{4!} e^c > 0$ . Thus  $f(x) - T_0^3 f(x) \geq 0$  which proves our claim.

**6.A.14.** In this case we have  $k = 2$  and hence we compute the error  $\frac{-x^3}{3(1+x)^3}$ .

**6.A.15.** We compute  $T_0^4 \sin(x) = x - \frac{x^3}{6}$ . In Sage the task can be solved by the block

```
f(x)=sin(x); F=plot(f(x), x, -pi, pi)
T4(x)=taylor(f(x), x, 0, 4); show(T4(x))
T=plot(T4(x), x, -pi, pi, color="black"); show(F+T)
```

This confirms the expression of  $T_0^4 \sin(x)$  and produces the graphs of  $f, T_0^4 f$ , which we present in the figure below (in this figure, with blue is coloured the graph of  $f$ )



Next we may approximate  $\sin 1^\circ$  by the expression  $\sin 1^\circ \approx \frac{\pi}{180} - \frac{\pi^3}{6 \cdot 180^3}$ .

**6.A.16.** The answer is  $1 - \frac{\pi^2}{10^2 \cdot 2} + \frac{\pi^4}{10^4 \cdot 4!}$ .

**6.A.18.** One can use the known formula  $\frac{1}{1+x} = \sum_{n=0}^{\infty} (-x)^n = \sum_{n=0}^{\infty} (-1)^n x^n$  corresponding to geometric series. By differentiating it, we obtain

$$-\frac{1}{(1+x)^2} = \left( \sum_{n=0}^{\infty} (-1)^n x^n \right)' = \sum_{n=1}^{\infty} (-1)^n n x^{n-1}.$$

Thus we see that

$$\frac{1}{(1+x)^2} = \sum_{n=1}^{\infty} (-1)^{n+1} n x^{n-1} = 1 - 2x + 3x^2 - 4x^3 + 5x^4 - 6x^5 + 7x^6 - 8x^7 + \dots$$

for all  $x \in (-1, 1)$ . Finally, to confirm this expression by Sage we use the same method as in 6.A.17, i.e.,

```
f(x)=1/(1+x)^2; tf=taylor(f, x, 0, 10); T=tf.power_series(QQ); show(T)
```

Sage's answer has the form

$$1 - 2x + 3x^2 - 4x^3 + 5x^4 - 6x^5 + 7x^6 - 8x^7 + 9x^8 - 10x^9 + 11x^{10} + O(x^{11}).$$

**6.A.19.** The first limit equals to  $-1/6$ , while the second limit takes the value  $-1$ .

**6.A.22.** An example is given by  $f(x) = x^4$ . The point  $x_0 = 0$  is a local minimum of  $f$  but  $f'(0) = 0$  and  $f''(0) = 0$ .

**6.A.26.** For the solution one can adopt the method presented in 6.A.25, which means that we should use appropriately the commands `find_local_maximum(f, a, b)` and `find_local_minimum(f, a, b)`. The implementation takes the form

```
f(x)=(1/8)*x^8-(3/5)*x^5-3*x+9
fp=plot(f(x), x, -2, 2, color="black", legend_label=r"$f(x)$")
show(fp)
show(find_local_maximum(f, -2, 2))
show(find_local_minimum(f, -2, 2))
g(x)=x^2-(1/6)*x^3
gp=plot(g(x), x, -2, 2, color="black", legend_label=r"$g(x)$")
show(gp)
show(find_local_maximum(g, -2, 2))
show(find_local_minimum(g, -2, 2))
h(x)=x/sqrt(x^2+4)
hp=plot(h(x), x, -2, 2, color="black", legend_label=r"$h(x)$")
show(hp)
show(find_local_maximum(h, -2, 2))
show(find_local_minimum(h, -2, 2))
k(x)=ln(x^3+8)
kp=plot(k(x), x, -2, 2, color="black", legend_label=r"$k(x)$")
show(kp)
show(find_local_maximum(k, -2, 2))
show(find_local_minimum(k, -2, 2))
```

Sage's answers look like as follows:

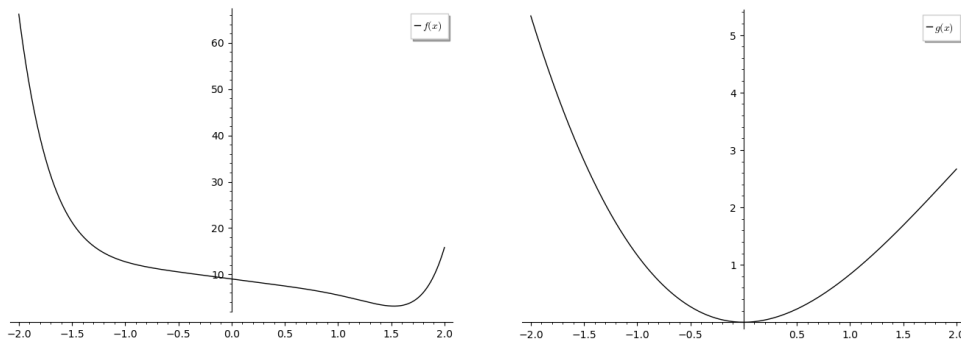
For  $f$  : (66.19999293835187, -1.9999999605494494) , (3.1326992998650347, 1.525960530898268) .

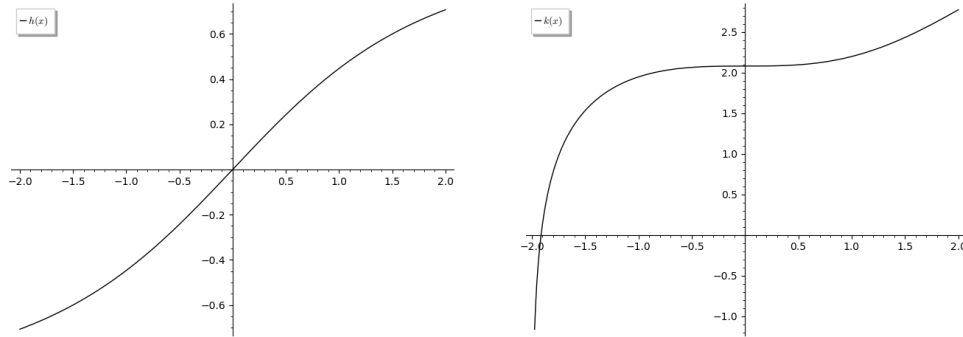
For  $g$  : (5.333333096630033, -1.9999999605494494) , (9.947977045494631  $\times 10^{-19}$ , 9.97395460544528e-10) .

For  $h$  : (0.7071067742126095, 1.9999999605494494) , (-0.7071067742126095, -1.9999999605494494) .

For  $k$  : (2.7725886926518686, 1.9999999605494494) , (-14.563311203591136, -1.9999999605494494) .

Here, recall that Sage first presents the maximal/minimum value  $f(x_0)$  at the extreme point  $x_0$ , and next the point  $x_0$  itself. It also produces the plots of the functions at hand, which are listed below.





**6.A.29.** The function  $f(x) = x^a$  is twice differentiable on  $(0, +\infty)$  with  $f'(x) = ax^{a-1}$  and  $f''(x) = a(a-1)x^{a-2}$ . By assumption we have  $a > 0$  and  $(a-1) < 0$ . It is also easy to see that  $x^{a-2} > 0$  for all  $x > 0$ . Thus  $f''(x) < 0$  for all  $x > 0$  which means that  $f$  is concave on  $(0, +\infty)$ .

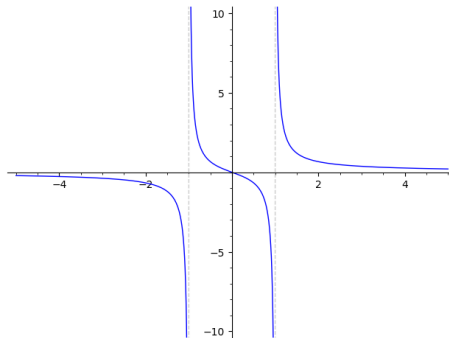
**6.A.30.** Based on the cell

```
f(x)=x/(x^2-1)
a=lim(f(x), x=1, dir="right")
b=lim(f(x), x=1, dir="left")
c=lim(f(x), x=-1, dir="right")
d=lim(f(x), x=-1, dir="left")
show(a, ', ', b, ', ', c, ', ', d)
```

we see that

$$a := \lim_{x \rightarrow 1^+} f(x) = +\infty, \quad b := \lim_{x \rightarrow 1^-} f(x) = -\infty, \quad c := \lim_{x \rightarrow -1^+} f(x) = +\infty, \quad d := \lim_{x \rightarrow -1^-} f(x) = -\infty.$$

Thus, the lines  $x = \pm 1$  are vertical asymptotes of  $f$ . It is also easy to see that the  $x$ -axis  $y = 0$  is the unique horizontal asymptote of  $f$  (this is because  $\lim_{x \rightarrow \infty} f(x) = 0 = \lim_{x \rightarrow -\infty} f(x)$ ). However, asymptotes with slope cannot exist, since  $f(x) = x/(x^2 - 1) = g(x)/h(x)$ , and the degree of  $g(x)$  is smaller than the degree of  $h(x)$ .



Recall that using Sage, to plot a rational function it is wise to restrict the  $y$ -values. On the other hand, Sage often sketches the vertical asymptotes via the option `detect_poles`, which we should appropriately include inside the command `plot`. In particular, the cell

```
p=plot(x/(x^2-1), x, -5, 5, ymin=-10, ymax=10,
detect_poles="show")
```

produces the figure posed above. However, one has the chance to set `detect_poles` to "True", (or "False"), to obtain only the graph of  $f$ . For an example with horizontal asymptotes see the [6.A.31](#).

**6.A.31.** (a) The given function is continuous and differentiable on  $\mathbb{R}$  as a fraction of differentiable functions. We see that its first derivative is everywhere positive,

$$f'(x) = \frac{2e^x}{(e^x + 1)^2} > 0, \quad x \in \mathbb{R}.$$

This implies that  $f$  is strictly increasing for all  $x \in \mathbb{R}$ . In case you want to verify this claim in Sage, use the cell

```
f(x) = (exp(x) - 1)/(exp(x) + 1); d1(x) = diff(f, x).factor(); bool(d1(x) > 0)
```



Now, since the domain of  $f$  is the whole real line,  $f$  cannot have vertical asymptotes. On the other hand, the lines  $y = \pm 1$  are horizontal asymptotes of  $f$  since

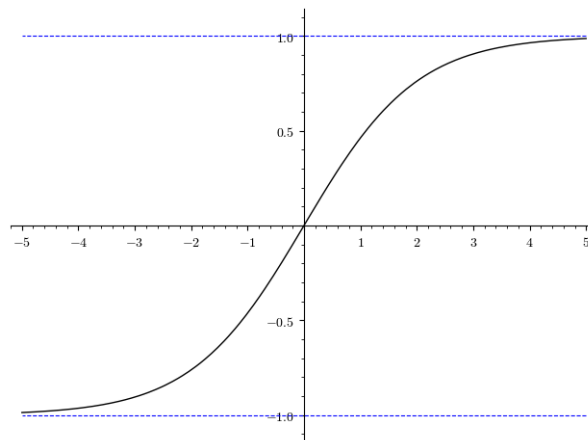
$$\lim_{x \rightarrow +\infty} \frac{e^x - 1}{e^x + 1} = \lim_{x \rightarrow +\infty} \frac{e^x}{e^x} = 1, \quad \lim_{x \rightarrow -\infty} \frac{e^x - 1}{e^x + 1} = \frac{0 - 1}{0 + 1} = -1.$$

Combining these conclusions one deduces that the range is the open interval  $(-1, 1)$ , see also the graph of  $f$  below.

It seems that Sage has not an “built-in” method for horizontal asymptotes. But we can proceed “manually” and introduce these lines via the command line. For instance, to sketch the graph of  $f$ , together with its asymptotes  $y = \pm 1$ , use the cell

```
f(x)=(exp(x)-1)/(exp(x)+1)
gf=plot(f, x, -5, 5, ymin=-1.1, ymax=1.1, color="black")
a1 = line([(5,1), (-5,1)], linestyle="--", thickness =0.8)
a2 = line([(5,-1), (-5,-1)], linestyle="--", thickness =0.8)
(gf+a1+a2).show(ticks=1,tick_formatter =1)
```

Executing this block, you will get the following figure



In this figure we have restricted the  $y$ -values, by the options  $ymin = -1.1$ , and  $ymax = 1.1$ , though not necessary (more experienced programmers could jump this step).

(b) We know that  $f$  is strictly increasing in  $\mathbb{R}$ , hence it does not admit (local/global) extremes. The second derivative of  $f$  is given by  $f''(x) = \frac{2(1-e^x)e^x}{(e^x+1)^3}$ , for all  $x \in \mathbb{R}$ . Thus,  $x = \ln(1) = 0$  is the unique solution of the equation  $f''(x) = 0$  and we have

- $f''(x) > 0$  for all  $x \in (-\infty, 0)$ ,
- $f''(x) < 0$  for all  $x \in (0, +\infty)$ .

Therefore,  $f$  is strictly convex for all  $x \in (-\infty, 0)$  and strictly concave for all  $x \in (0, +\infty)$ , and since  $f$  is continuous on 0 and its graph there changes from convex to concave,  $x = 0$  is the unique inflection point of  $f$  (with value  $f(0) = 0$ ).

Some of the previous conclusions occur really easy in Sage, as well. For instance, by typing

```
f(x)=(exp(x)-1)/(exp(x)+1)
d2(x)=diff(f, x, 2); show(solve(d2(x)==0, x))
```

we confirm the inflection point of  $f$ .

**6.A.34.** By assumption,  $f$  is differentiable on  $\mathbb{R}$ . Hence, if  $a \in \mathbb{R}$  is a stationary point of  $f$  we have  $f'(a) = 0$ . On the other hand, a differentiation of the defining equation gives

$$2f(x)f'(x) - f(x) - xf'(x) + 4x = 0, \quad x \in \mathbb{R}. \quad (*)$$

By replacing  $x = a$  in this equation, we find that  $f(a) = 4a$ . Combining this with the defining equation, the replacement  $x = a$  gives  $f^2(a) - af(a) + 2a^2 - 7 = 0$ , that is,  $(4a)^2 - 4a^2 + 2a^2 - 7 = 0$ , which is equivalent to  $14a^2 = 7$ . Thus  $a^2 = \frac{1}{2}$ , i.e.,  $a = \pm \frac{\sqrt{2}}{2}$ .

Next,  $f$  is twice differentiable. Hence, if  $x = b \in \mathbb{R}$  is an inflection point of  $f$  we should have  $f''(b) = 0$ . Now, taking the derivative of  $(*)$ , we see that

$$2(f'(x))^2 + 2f(x)f''(x) - 2f'(x) - xf''(x) + 4 = 0, \quad x \in \mathbb{R}.$$

For  $x = b$  this relation reduces to  $(f'(b))^2 - f'(b) + 2 = 0$ , which has negative discriminant,  $\Delta = -7 < 0$ . This implies that  $f$  does not admit inflection points.

**6.A.35.** We have  $f'(x) = 2x e^{x^2}$ ,  $g'(x) = \frac{2x}{x^2+1}$ , and  $k'(x) = \frac{e^x}{e^{2x}+1}$ , for all  $x \in \mathbb{R}$ . Thus, we get  $df(x) = (2x e^{x^2}) dx$ ,  $dg(x) = \frac{2x}{x^2+1} dx$ , and  $dk(x) = \frac{e^x}{e^{2x}+1} dx$ , respectively. An implementation of these differentials in Sage, essentially relies on the command `diff` (or `derivative`). For instance, a cell based on this method and which includes  $dx$  as a symbolic variable has the form

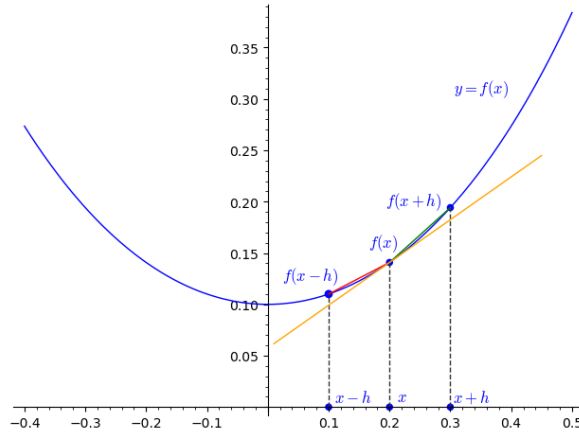
```
var("x, dx")
f(x)=e**(x**2); f1(x)=diff(f, x)
df(x)=f1*dx #declare the differential of f
g(x)=ln(x**2+1); g1(x)=diff(g, x)
dg(x)=g1*dx #declare the differential of g
k(x)=arctan(e^x); k1(x)=diff(k, x)
dk(x)=k1*dx #declare the differential of k
show(df(x), " ", dg(x), " ", dk(x))
```

Chek yourself Sage's output.

**6.A.41.** First we need to introduce  $f$  and use the command `plot` to sketch it. We also need to indicate the nodes  $x$ ,  $x+h$ ,  $f(x)$  and  $f(x+h)$ , and draw the slopes corresponding to forward/backward differences, together with the tangent line. Hence, we may appropriately use the commands `point` and `line` (recall that these are build-in functions in Sage for the construction of points and lines, respectively). To keep our code simple, we agree to distinguish the slopes via colours. So, let us use green and red colours to illustrate the forward/backward difference slopes, and orange colour for the tangent line of  $f$  trough  $x$ . We can now encode the solution, by the following block:

```
f(x)=exp(x^(2))-0.9
p=plot(f(x), x, -0.4, 0.5, ticks=[0.1,None])
p+=point((0.2, f(0.2)), size=30)
p+=point((0.3, f(0.3)), size=30)
p+=point((0.2, 0), size=30)
p+=point((0.3, 0), size=30)
p+=line([(0.2, f(0.2)), (0.3, f(0.3))], color="green")
p+=line([(0.2, 0), (0.2, f(0.2))],
linestyle="--", rgbcolor=(0.2,0.2,0.2))
p+=line([(0.3, 0), (0.3, f(0.3))],
linestyle="--", rgbcolor=(0.2,0.2,0.2))
p+=point((0.1, 0), size=30)
p+=point((0.1, f(0.1)), size=40)
p+=line([(0.1, f(0.1)), (0.2, f(0.2))], color="red")
p+=line([(0.1, 0), (0.1, f(0.1))],
linestyle="--", rgbcolor=(0.2,0.2,0.2))
f1=diff(f(x), x)(x=0.2); l1(x)=f(0.2)+f1*(x-0.2)
p+=plot(l1(x), x, 0.01, 0.45, color="orange")
p+=text(r"$x$", (0.22, 0.01), fontsize=12)
p+=text(r"$x+h$", (0.335, 0.01), fontsize=12)
p+=text(r"$x-h$", (0.14, 0.01), fontsize=12)
p+=text(r"$f(x)$", (0.19, 0.16), fontsize=12)
p+=text(r"$f(x+h)$", (0.24, 0.20), fontsize=12)
p+=text(r"$f(x-h)$", (0.07, 0.128), fontsize=12)
p+=text(r"$y=f(x)$", (0.35, 0.31), fontsize=12)
show(p)
```

This block has as output the following figure:



**6.A.42.** The central difference approximation of the value  $f'(0.2) = 0.4163243$  with step  $h = 0.01$  is given by

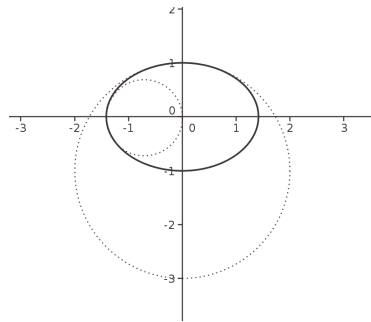
$$\frac{f(0.2 + 0.01) - f(0.2 - 0.01)}{0.02} \approx 0.4163671.$$

This proves the claim, comparing with the results in 6.A.40.

**6.A.43.** The axes of the given ellipse are the coordinate axes  $x$  and  $y$ , and its vertices are the points

$$P = [\sqrt{2}, 0], Q = [-\sqrt{2}, 0], R = [0, 1], S = [0, -1],$$

see also the given figure below.



Let us first compute the curvature at  $R$ . We can consider the coordinate  $y$  as a function of  $x$  (determined uniquely in a neighbourhood of  $R$ ). Thus by differentiating the equation of the ellipse with respect to  $x$  we get  $2x + 4yy' = 0$ . Hence  $y' = -\frac{x}{2y}$ . Differentiating this equation once more with respect to  $x$  we obtain the second derivative,

$$y'' = -\frac{1}{2} \left( \frac{1}{y} - \frac{xy'}{y^2} \right).$$

At the point  $R$ , we see that  $y'(R) = 0$  and  $y''(R) = -\frac{1}{2}$ . According to 6.1.13, the radius of the osculation circle will be given by the expression

$$\frac{(1 + (y')^2)^{\frac{3}{2}}}{(y'')^2}$$

evaluated at the point in question. This gives the values  $\pm 2$  (the sign tells us the circle will be “below” the graph of the function). Now, the ideas in 6.1.13 and 6.1.16 imply that the centre will be in the direction opposite to the normal line of this curve, i.e., on the  $y$ -axis. Since the radius is 2, the centre will be at the point  $[0, 1 - 2] = [0, -1]$ . In total, the equation of the osculation circle of the given ellipse at  $R$  will be  $x^2 + (y + 1)^2 = 4$ . Analogously, we can determine the equation of the osculation circle at the point  $S$ , where we get the equation  $x^2 + (y - 1)^2 = 4$ . Finally, the curvatures of the ellipse at these points equals to  $\frac{1}{2}$  (the absolute value of the curvature of the graph of the function).

For determining the osculation circle at the point  $[\sqrt{2}, 0]$ , we’ll consider the equation of the ellipse as a formula for the variable  $x$  depending on the variable  $y$ , i.e.,  $x$  as a function of  $y$ , determined uniquely, so we cannot use the previous procedure - technically it would end up by dividing by zero). Then, by differentiation we obtain:  $2xx' + 4y = 0$ , thus  $x' = -2\frac{y}{x}$ , and  $x'' = -2\left(\frac{1}{x} - \frac{yx'}{x^2}\right)$ . Hence at point  $[\sqrt{2}, 0]$ , we have  $x' = 0$  and  $x'' = -\sqrt{2}$  and the radius of the circle of osculation is  $\rho = -\frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$  according to 6.1.13. The normal line is heading to  $-\infty$  along the  $x$  axis at point  $[\sqrt{2}, 0]$ , thus the center

of the osculation circle will be on the  $x$  axis on the other side at distance  $\frac{\sqrt{2}}{2}$ , hence at the point  $[\sqrt{2} - \frac{\sqrt{2}}{2}, 0] = [\frac{\sqrt{2}}{2}, 0]$ . In total, the equation of the circle of osculation at vertice  $[\sqrt{2}, 0]$  will be  $(x - \frac{\sqrt{2}}{2})^2 + y^2 = \frac{1}{2}$ . The curvature at both of these vertices equals  $\sqrt{2}$ .

Note that the vertices of an ellipse (more generally the vertices of a closed smooth curve in plane) can be defined as the points at which the function of curvature has an extreme. The ellipse having four vertices isn't a coincidence. The so called "four vertices theorem" states that a closed curve of the class  $C^3$  has at least four vertices. A curve of the class  $C^3$  is locally given parametrically by points  $[f(t), g(t)] \in \mathbb{R}^2$ ,  $t \in (a, b) \subset \mathbb{R}$ , where  $f$  and  $g$  are functions of the class  $C^3(\mathbb{R})$ . Thus the curvature of the ellipse at its any point is between its curvatures at its vertices, i.e. between  $\frac{1}{2}$  and  $\sqrt{2}$ .

**6.A.44.** (a) Since the functions  $x = x(t)$  and  $y = y(t)$  are differentiable, by the chain rule one has

$$y'(t) = \frac{dy}{dt} = \frac{dy}{dx} \frac{dx}{dt} \implies \frac{dy}{dx} = \frac{dy/dt}{dx/dt},$$

assuming that  $x'(t) = 0$  for all  $t \in I$ .

(b) For  $\alpha(t)$  we have

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{6t^2}{1} = 6t^2.$$

Thus  $\frac{dy}{dx}|_{t=0} = 0$  and  $\frac{dy}{dx}|_{t=2} = 24$ . Alternatively, we have  $x = 1 + t$ , thus  $t = x - 1$  and hence  $y = 2(x - 1)^3$ . Thus  $y'(x) = \frac{dy}{dx} = 6(x - 1)^2 = 6t^2$ .

Next, for  $\beta(t)$  we have

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = -\frac{(1+t^2)^2}{2t(1+t)}. \quad (b)$$

However in this case we don't have  $x'(t) \neq 0$  for all  $t \in I$ , in particular, the derivative  $\frac{dy}{dx}$  is not defined at  $t = 0$ . However, at  $t = 2$  one gets  $\frac{dy}{dx}|_{t=2} = -25/12$ . To confirm the situation in an alternative way, we need to eliminate  $t$  which can be done by the first parametric equation  $x = 1/(1+t^2)$ . We get  $t = \sqrt{\frac{1-x}{x}}$  (notice that  $x = 1$  for  $t = 0$  and  $x = 1/5$  for  $t = 2$ ). Thus,  $y(x) = \ln\left(1 + \sqrt{\frac{1-x}{x}}\right)$  and now one can directly differentiate  $y$  with respect to  $x$ . Lets do this quickly in Sage:

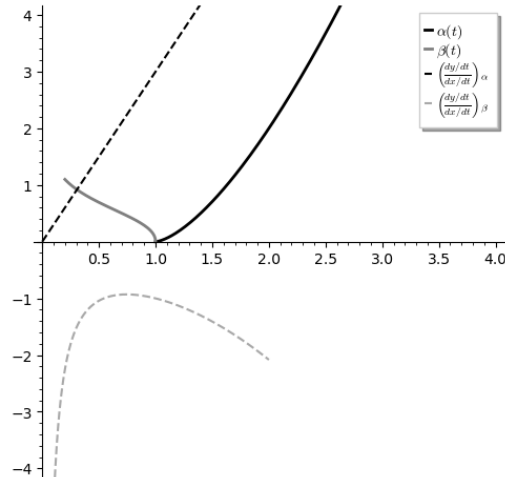
```
var("t")
f(x)=ln(1+sqrt((1-x)/x))
show(diff(f, x))
show(diff(f, x)(x=1/5))
show(diff(f, x)(x=1))
show(diff(f(x), x)(x=1/(1+t^2)).full_simplify())
```

The very first show command returns the derivative  $y'(x)$  (in terms of  $x$ ), the second the value  $y'(1/5) = -25/12$  while the third one returns an error, since there  $y'(x)$  is not defined. Finally, the last command confirms the expression given in (b).

(c) Let us now pose the code for solving this task:

```
var("t")
F1(t)=(diff(2*t^3, t)/diff(t^2+1, t)).factor(); show(F1)
F2(t)=(diff(ln(1+t), t)/diff(1/(1+t^2), t)).factor(); show(F2)
p=parametric_plot((t^2+1, 2*t^3), (t,0,2), legend_label=r"\alpha(t)", color="black", thickness=2)
p+=parametric_plot((1/(1+t^2), ln(1+t)), (t,0,2), legend_label=r"\beta(t)", color="grey", thickness=2)
p+=plot(F1(t), t, 0, 2, color="black", linestyle="--", thickness=1.5,
legend_label=r"\left(\frac{dy}{dt}\right)_{\alpha}")
p+=plot(F2(t), t, 0, 2, color="darkgrey", linestyle="--", thickness=1.5,
legend_label=r"\left(\frac{dy}{dt}\right)_{\beta}")
p.show(ymin=-4, ymax=4, xmax=4, aspect_ratio=1/2)
```

This block confirms the given expressions of the derivatives  $\frac{dy}{dx}$  and produces the graphs of the curves  $\alpha, \beta$ , together with the graphs of the corresponding derivatives.



**6.A.46.** Let us present the block with some comments within:

```
var("t, s, r")
x(t)=t-sin(t); y(t)=1-cos(t)
show(diff(y, t)/diff(x, t)) # declare the slope of the cycloid at general point
y0=y(2*pi/3);show(y0) # declare the y-coordinate of the point P
x0=x(2*pi/3);show(x0) # declare the x-coordinate of the point P
k=diff(y, t)(t=2*pi/3)/(diff(x, t)(t=2*pi/3)); show(k) # declare the slope of the cycloid at 2pi/3
l(s)=y0+k*s-k*x0; show(l) #declare the tangent line passing through P
y00=y(4*pi/3); show(y00) # declare the y-coordinate of the point Q
x00=x(4*pi/3); show(x00) # declare the x-coordinate of the point Q
kk=diff(y, t)(t=4*pi/3)/(diff(x, t)(t=4*pi/3)); show(kk) # declare the slope of the cycloid at 4pi/3
L(r)=y00+kk*r-kk*x00; show(L) #declare the tangent line passing through Q
show(solve(l(r)-L(r)==0, r)) #declare the intersection point of the 2 tangent lines
p=parametric_plot((x(t), y(t)), (t,0,2*pi), color="black", thickness=2)
p+=plot(l(s), -0.5, 4, color="darkgrey")
p+=plot(L(r), 2, 6, color="grey")
p+=point([x(2*pi/3), y(2*pi/3)], size=30) # the point P
p+=point([x(4*pi/3), y(4*pi/3)], size=30) # the point Q
p+=point([pi, l(pi)], size=30) # the intersection point R
p+=text(r"$P$", (x0, y0+0.2), fontsize=14, rgbcolor=(0.1,0.2,0.5))
p+=text(r"$Q$", (x00, y00+0.2), fontsize=14, rgbcolor=(0.1,0.2,0.5))
p+=text(r"$R$", (pi, l(pi)+0.2), fontsize=14, rgbcolor=(0.1,0.2,0.5))
p.show(gridlines="true")
```

**6.B.2.** By definition, we have  $F'(x) = f(x)$  for all  $x \in \mathbb{R}$ . Combining this with the relation  $4F(x) = f(x)$  we get  $F'(x) = 4F(x)$ , that is,  $\frac{F'(x)}{F(x)} = 4$  (since  $F(x) = (1/4)f(x) \neq 0$  for all  $x \in \mathbb{R}$ ). An integration then gives

$$\int \frac{F'(x)}{F(x)} dx = \int 4 dx \iff \ln |F(x)| = 4x + c \iff |F(x)| = e^{4x+c},$$

for some constant  $c$ . Thus  $F(x) = e^{4x+c}$  or  $F(x) = -e^{4x+c}$ . Because  $F(4) = \frac{f(4)}{4} = 1 > 0$  the second case is omitted. In particular, the relation  $F(4) = 1$  gives  $e^{16+c} = 1 = e^0$ , that is,  $16 + c = 0$  or equivalently  $c = -16$ . Thus  $F(x) = e^{4(x-4)}$  and  $f(x) = F'(x) = 4e^{4(x-4)}$ .

**6.B.4.** One can use the identities  $\sin^2(x) = (1 - \cos(2x))/2$  and  $\cos^2(x) = (1 + \cos(2x))/2$ . In particular, we see that

$$\begin{aligned} A &= \int \left( \frac{1 - \cos(2x)}{2} \right) \left( \frac{1 + \cos(2x)}{2} \right) dx = \int \frac{1 - \cos^2(2x)}{4} dx = \frac{1}{4} \int dx - \frac{1}{4} \int \cos^2(2x) dx \\ &= \frac{x}{4} - \frac{1}{4} \int \left( \frac{1 + \cos(4x)}{2} \right) dx = \frac{x}{4} - \frac{1}{8} \int dx - \frac{1}{8} \int \cos(4x) dx = \frac{x}{8} - \frac{\sin(4x)}{32} + C. \end{aligned}$$

**6.B.7.** (a) We have  $(\sin(x))' = \cos(x)$  and hence

$$K = \int x(\sin(x))' dx = x \sin(x) - \int (x)' \sin(x) dx = x \sin(x) + \cos(x) + C, \quad C \in \mathbb{R}.$$

(b) Recall that  $(x^2 - x)' = 2x - 1$  and hence

$$M = \int (x^2 - x)' \ln(x) dx = (x^2 - x) \ln(x) - \int (x^2 - x)(\ln(x))' dx = (x^2 - x) \ln(x) - \int (x - 1) dx = (x^2 - x) \ln(x) + x - \frac{x^2}{2} + C.$$

(c) In this case we will use the relations  $(-\frac{1}{\beta} \cos(\beta x))' = \sin(\beta x)$  and  $(\frac{1}{\beta} \sin(\beta x))' = \cos(\beta x)$ , for all  $x \in \mathbb{R}$ . Thus

$$N = \int e^x \sin(\beta x) dx = \int e^x (-\frac{1}{\beta} \cos(\beta x))' dx = -\frac{1}{\beta} e^x \cos(\beta x) + \frac{1}{\beta} \nu, \quad (*)$$

where  $\nu = \int e^x \cos(\beta x) dx$ . In the same vein we see that

$$\nu = \int e^x (\frac{1}{\beta} \sin(\beta x))' dx = \frac{1}{\beta} e^x \sin(\beta x) - \frac{1}{\beta} N. \quad (**)$$

Therefore, a combination of (\*) and (\*\*) gives

$$N + \frac{1}{\beta^2} N = -\frac{1}{\beta} e^x \cos(\beta x) + \frac{1}{\beta^2} e^x \sin(\beta x) + C,$$

that is,  $N = \frac{\beta}{\beta^2 + 1} e^x (\frac{1}{\beta} \sin(\beta x) - \cos(\beta x)) + C$ ,  $C \in \mathbb{R}$ .

A word of caution: One should remember to identify multiples of  $C$  with  $C$  itself, as we did in cases (b) and (d), for example. Moreover, we could even write  $N(x) = \frac{\beta}{\beta^2 + 1} e^x (\frac{1}{\beta} \sin(\beta x) - \cos(\beta x)) + C$ ,  $x \in \mathbb{R}$ , and similarly for the previous cases (we avoid this to save some space).

**6.B.8.** All the solutions can be obtained by applying the rule  $\int F(x)G'(x) dx = F(x)G(x) - \int F'(x)G(x) dx$  (integration by parts). Notice the first non-trivial part in this rule is the function  $G(x)$ , which we should “guess” using its first derivative  $G'(x)$ . Next we should be able to compute the integral  $\int F'(x)G(x) dx$ .

(a) We have  $F(x) = x$ ,  $F'(x) = 1$ ,  $G'(x) = \frac{1}{\cos^2(x)}$  and hence  $G(x) = \tan(x)$ , such that  $(\tan(x))' = \frac{1}{\cos^2(x)}$ . Therefore, an application of the preceding rule gives

$$\begin{aligned} \int \frac{x}{\cos^2(x)} dx &= \int x(\tan(x))' dx = x \tan(x) - \int \tan(x) dx = x \tan(x) + \int \frac{-\sin(x)}{\cos(x)} dx \\ &= x \tan(x) + \int \frac{(\cos(x))'}{\cos(x)} dx = x \tan(x) + \ln |\cos(x)| + C. \end{aligned}$$

Above, in the final step we applied the identity  $\int \frac{f'(x)}{f(x)} dx = \ln |f(x)| + C$ , for some constant  $C$  (see 6.2.3).

(b) In this case we have  $F(x) = x^2$ ,  $F'(x) = 2x$ ,  $G'(x) = e^{-3x}$  and hence  $G(x) = -\frac{1}{3} e^{-3x}$ . Therefore, at a first step we obtain

$$\int x^2 e^{-3x} dx = \int x^2 (-\frac{1}{3} e^{-3x})' dx = -\frac{1}{3} x^2 e^{-3x} + \frac{2}{3} \int x e^{-3x} dx.$$

Next we see that

$$\int x e^{-3x} dx = \int x (-\frac{1}{3} e^{-3x})' dx = -\frac{1}{3} x e^{-3x} + \frac{1}{3} \int e^{-3x} dx = -\frac{1}{3} x e^{-3x} - \frac{1}{9} e^{-3x} + C,$$

for some constant  $C \in \mathbb{R}$ , where in the final step we used the relation  $\int e^{ax} dx = \frac{e^{ax}}{a} + C$  (see 6.2.3). A combination of these two relations gives

$$\int x^2 e^{-3x} dx = -\frac{1}{3} x^2 e^{-3x} + \frac{2}{3} \left( -\frac{1}{3} x e^{-3x} - \frac{1}{9} e^{-3x} \right) + C = -\frac{1}{3} e^{-3x} \left( x^2 + \frac{2}{3} x + \frac{2}{9} \right) + C,$$

for some constant  $C \in \mathbb{R}$ .

(c) In this case we have  $F(x) = \cos(x)$ ,  $F'(x) = -\sin(x)$ ,  $G'(x) = \cos(x)$  and hence  $G(x) = \sin(x)$ . If we set for convenience  $M = \int \cos^2(x) dx$ , we thus have

$$\begin{aligned} M &= \int \cos^2(x) dx = \int \cos(x) \cos(x) dx = \int \cos(x) (\sin(x))' dx = \cos(x) \sin(x) + \int \sin^2(x) dx \\ &= \cos(x) \sin(x) + \int (1 - \cos^2(x)) dx = \cos(x) \sin(x) + \int 1 dx - \int \cos^2(x) dx = \cos(x) \sin(x) + x - M + C, \end{aligned}$$

that is,  $2M = \cos(x)\sin(x) + x + C$  or equivalently,  $M = \frac{1}{2}(\cos(x)\sin(x) + x) + C$ , for some constant  $C \in \mathbb{R}$ . We emphasise that usually suitable simplifications or substitutions, lead to the desired result faster than integration by parts. For instance, by using the identity

$$\cos^2(x) = \frac{1}{2}(1 + \cos(2x)), \quad x \in \mathbb{R}$$

we easily obtain

$$\int \cos^2(x) dx = \int \frac{1}{2} dx + \int \frac{1}{2} \cos(2x) dx = \frac{x}{2} + \frac{\sin(2x)}{4} + C = \frac{x}{2} + \frac{2\sin(x)\cos(x)}{4} + C = \frac{1}{2}(x + \sin(x) \cdot \cos(x)) + C,$$

for some constant  $C \in \mathbb{R}$ .

**6.B.10.** Set  $t = x^3 + 4$  such that  $dt = 3x^2 dx$ . Thus  $dx = \frac{dt}{3x^2}$  and

$$\int e^{x^3+4} x^2 dx = \frac{1}{3} \int e^t dt = \frac{e^t}{3} + C = \frac{e^{x^3+4}}{3} + C,$$

for some constant  $C$ . In Sage one can type

```
show(integral((e^(x^3+4))*x^2, x))
```

**6.B.11.** The main idea of the so called *first substitution method* is writing the integral in the form of

$$\int f(\varphi(x)) \varphi'(x) dx, \quad (\star)$$

for certain functions  $f$  and  $\varphi$ . Then the substitution  $y = \varphi(x)$  gives that  $dy = \varphi'(x) dx$ , and the integral above reads as

$$\int f(y) dy.$$

Based on this idea let us set  $y = \cos(x)$ , such that  $dy = -\sin(x) dx$ . Then we obtain

$$\int \cos^5(x) \sin(x) dx = - \int \cos^5(x) (-\sin(x)) dx = - \int y^5 dy = -\frac{y^6}{6} + C = -\frac{\cos^6(x)}{6} + C,$$

for some arbitrary constant  $C \in \mathbb{R}$ .

**6.B.12.** We will use the so called *second substitution method*, which means a reduction of  $\int f(y) dy$  to the form  $(\star)$  presented in the previous task [6.B.11](#), for  $y = \varphi(x)$ . In particular, in our example we want to determine the primitive function of function  $f(x) = \tan^4(x)$ . Thus it is sensible to consider the substitution  $u = \tan(x)$ , and hence  $x = \arctan(u)$ . This gives  $dx = \frac{du}{1+u^2}$ , and hence we get

$$\int \frac{\sin^4(x)}{\cos^4(x)} dx = \int \frac{u^4}{1+u^2} du = \int u^2 - 1 + \frac{1}{u^2+1} du = \frac{u^3}{3} - u + \arctan(u) + C = \frac{\tan^3(x)}{3} - \tan(x) + x + C,$$

for some constant  $C \in \mathbb{R}$ .

**6.B.13.** Let us include the solutions in one cell:

```
show(integral((cos(x))^5*sin(x), x)); show(integral((sin(x))^4/((cos(x))^4), x))
```

Check yourself that this verifies the given answers in [6.B.11](#) and [6.B.12](#), respectively.

**6.B.14.** This is based on appropriate substitution, in particular

$$\begin{aligned} \int \cos^5(x) \sin^2(x) dx &= \int (\cos^2(x))^2 \sin^2(x) \cos(x) dx = \int ((1 - \sin^2(x))^2 \sin^2(x) \cos(x) dx \\ &\stackrel{t=\sin(x)}{dt=\cos(x) dt} \int (1 - t^2)^2 t^2 dt = \int (1 - 2t^2 + t^4) t^2 dt = \int (t^6 - 2t^4 + t^2) dt \\ &= \frac{t^7}{7} - 2\frac{t^5}{5} + \frac{t^3}{3} + C = \frac{\sin^7(x)}{7} - 2\frac{\sin^5(x)}{5} + \frac{\sin^3(x)}{3} + C, \end{aligned}$$

for some constant  $C$ .

**6.B.15.** (a) The substitution method leads to the integral

$$\int x^3 e^{-x^2} dx = \left| \begin{array}{l} t = -x^2 \\ dt = -2x dx \end{array} \right| = \frac{1}{2} \int t e^t dt,$$

which can be easily computed by integrating by parts, yielding

$$\frac{1}{2} \int t e^t dt = \left| \begin{array}{l} F(t) = t \\ G'(t) = e^t \end{array} \right| \left| \begin{array}{l} F'(t) = 1 \\ G(t) = e^t \end{array} \right| = \frac{1}{2} t e^t - \frac{1}{2} \int e^t dt = \frac{1}{2} t e^t - \frac{1}{2} e^t + C = -\frac{1}{2} e^{-x^2} (x^2 + 1) + C, \quad C \in \mathbb{R}.$$

(b) Similarly, we obtain

$$\begin{aligned} \int x \arcsin(x^2) dx &= \left| \begin{array}{l} t = x^2 \\ dt = 2x dx \end{array} \right| = \frac{1}{2} \int \arcsin(t) dt = \left| \begin{array}{l} F(t) = \arcsin(t) \\ G'(t) = 1 \end{array} \right| \left| \begin{array}{l} F'(t) = \frac{1}{\sqrt{1-t^2}} \\ G(t) = t \end{array} \right| \\ &= \frac{1}{2} t \arcsin(t) - \frac{1}{2} \int \frac{t}{\sqrt{1-t^2}} dt = \left| \begin{array}{l} u = 1-t^2 \\ du = -2t dt \end{array} \right| = \frac{1}{2} t \arcsin(t) + \frac{1}{4} \int \frac{du}{\sqrt{u}} \\ &= \frac{1}{2} t \arcsin(t) + \frac{1}{2} \sqrt{u} + C = \frac{1}{2} t \arcsin(t) + \frac{1}{2} \sqrt{1-t^2} + C \\ &= \frac{1}{2} x^2 \arcsin(x^2) + \frac{1}{2} \sqrt{1-x^4} + C, \quad C \in \mathbb{R}. \end{aligned}$$

(c) In this case let us first use the substitution  $y = \sqrt{x}$  to get rid of the root from the argument of the exponential function. This leads to the integral

$$\int e^{\sqrt{x}} dx = \left| \begin{array}{l} y^2 = x \\ 2y dy = dx \end{array} \right| = 2 \int y e^y dy.$$

Based now on integration by parts, we obtain

$$\int y e^y dy = \left| \begin{array}{l} F(y) = y \\ G'(y) = e^y \end{array} \right| \left| \begin{array}{l} F'(y) = 1 \\ G(y) = e^y \end{array} \right| = y e^y - \int e^y dy = y e^y - e^y + C, \quad C \in \mathbb{R}.$$

In total this gives

$$\int e^{\sqrt{x}} dx = 2y e^y - 2e^y + C = 2e^{\sqrt{x}} (\sqrt{x} - 1) + C, \quad C \in \mathbb{R}.$$

**6.B.17.** Integration by parts gives

$$\begin{aligned} \int \sin^n(x) dx &= \int \sin^{n-1}(x) \sin(x) dx = \int \sin^{n-1}(x) (-\cos(x))' dx \\ &= -\sin^{n-1}(x) \cos(x) + (n-1) \int \sin^{n-2}(x) \cos^2(x) dx \\ &= -\sin^{n-1}(x) \cos(x) + (n-1) \int \sin^{n-2}(x) (1 - \sin^2(x)) dx \\ &= -\sin^{n-1}(x) \cos(x) + (n-1) \int \sin^{n-2}(x) dx - (n-1) \int \sin^n(x) dx \\ &= -\sin^{n-1}(x) \cos(x) + (n-1) I_{n-2} - (n-1) I_n. \end{aligned}$$

Thus for any positive integer  $n$  we get

$$n I_n = -\sin^{n-1}(x) \cos(x) + (n-1) I_{n-2} \iff I_n = -\frac{1}{n} \sin^{n-1}(x) \cos(x) + \frac{n-1}{n} I_{n-2}.$$

Now, we have  $I_0 = x$  and  $I_1 = -\cos(x)$ . Thus based on the recurrence relation and using the identity  $\sin(2x) = 2 \sin(x) \cos(x)$ , we get

$$I_2 = \int \sin^2(x) dx = -\frac{1}{2} \sin(x) \cos(x) + \frac{1}{2} I_0 = -\frac{1}{2} \sin(x) \cos(x) + \frac{1}{2} x = -\frac{1}{4} \sin(2x) + \frac{1}{2} x.$$

Similarly,

$$\begin{aligned} I_3 &= \int \sin^3(x) dx = -\frac{1}{3} \sin^2(x) \cos(x) + \frac{2}{3} I_1 = -\frac{1}{3} \sin^2(x) \cos(x) - \frac{2}{3} \cos(x) = -\frac{1}{3} (1 - \cos^2(x)) \cos(x) - \frac{2}{3} \cos(x) \\ &= \frac{1}{3} \cos^3(x) - \cos(x). \end{aligned}$$

**6.B.18.** We can solve this by substitution, which we will encode in a more “compact” but less informative way, as follows:

$$\int \frac{6}{x-2} dx = \left| \begin{array}{l} y = x-2 \\ dy = dx \end{array} \right| = \int \frac{6}{y} dy = 6 \ln |y| + C_1 = 6 \ln |x-2| + C_1, \quad C_1 \in \mathbb{R}.$$

Similarly, we have

$$\int \frac{6}{(x+4)^3} dx = \left| \begin{array}{l} y = x+4 \\ dy = dx \end{array} \right| = \int \frac{6}{y^3} dy = \frac{6}{-2y^2} + C_2 = -\frac{3}{(x+4)^2} + C_2, \quad C_2 \in \mathbb{R}.$$

**6.B.19.** Setting  $w = ax + b$  we get  $dw = a dx$ , that is,  $dx = \frac{dw}{a}$ . Hence



$$\int \frac{dx}{(ax+b)^n} = \int \frac{dw}{aw^n} = \frac{1}{a} \int w^{-n} dw = \frac{1}{a} \left( \frac{w^{-n+1}}{-n+1} \right) + C = -\frac{1}{a(n-1)w^{n-1}} + C,$$

and the result follows.

Next observe that  $9x^2 + 6x + 1 = (3x + 1)^2$ . Thus  $\int \frac{dx}{9x^2 + 6x + 1} = \int \frac{dx}{(3x + 1)^2}$ , and by applying the relation that we proved in (a), it follows that  $\int \frac{dx}{(3x + 1)^2} = -\frac{1}{3(3x + 1)} + C$ , for some constant  $C$ .

**6.B.20.** Recall that  $P(x) = ax^2 + bx + c = a \left(x + \frac{b}{2a}\right)^2 - \frac{\Delta}{4a}$ , and since  $\Delta = 0$  we get  $P(x) = a \left(x + \frac{b}{2a}\right)^2$ . Therefore,  $\int \frac{dx}{P(x)} = \int \frac{dx}{a \left(x + \frac{b}{2a}\right)^2}$ . Let us set  $u = x + \frac{b}{2a}$ , such that  $du = dx$ . Then we see that

$$\int \frac{dx}{a \left(x + \frac{b}{2a}\right)^2} = \frac{1}{a} \int \frac{du}{u^2} = \frac{1}{a} \int u^{-2} du = -\frac{1}{au} + C = -\frac{1}{a \left(x + \frac{b}{2a}\right)} + C = -\frac{2}{2ax + b} + C,$$

for some constant  $C$ . We leave the other confirmation to the reader.

**6.B.27.** By the discussion in 6.2.7 we can assume that

$$Q(x) = \frac{2x^4 + 2x^2 - 5x + 1}{x(x^2 - x + 1)^2} = \frac{A}{x} + \frac{B_1x + C_1}{x^2 - x + 1} + \frac{B_2x + C_2}{(x^2 - x + 1)^2},$$

for some  $A, B_1, C_1, B_2, C_2$ , to be specified. By equating coefficients and solving the corresponding system one gets  $A = 1, B_1 = 1, C_1 = 3, B_2 = 1$  and  $C_2 = -6$ , and thus

$$Q(x) = \frac{1}{x} + \frac{x + 3}{x^2 - x + 1} + \frac{x - 6}{(x^2 - x + 1)^2}.$$

Recall that you can use Sage to get a quick confirmation of the obtained fractional decomposition of  $Q$ , just by typing

```
Q(x)=(2*x^4+2*x^2-5*x+1)/(x*(x^2-x+1)^2); show(Q.partial_fraction())
```

The fractional decomposition of  $Q$  implies that the integral in question splits in three parts:

$$\int Q(x) dx = \int \frac{2x^4 + 2x^2 - 5x + 1}{x(x^2 - x + 1)^2} dx = \int \frac{1}{x} dx + \int \frac{x + 3}{x^2 - x + 1} dx + \int \frac{x - 6}{(x^2 - x + 1)^2} dx. \quad (\star)$$

Let us focus on the last two integrals which are more difficult. To simplify our notation, we will compute them up to a constant (as Sage does). We begin with  $\Lambda := \int \frac{x + 3}{x^2 - x + 1} dx$ . Since  $x^2 - x + 1$  has complex roots one can apply a method similar to those presented in 6.B.22. Hence we get

$$\Lambda = \int \frac{x + 3}{x^2 - x + 1} dx = \frac{1}{2} \int \frac{2x - 1}{x^2 - x + 1} dx + \frac{7}{2} \int \frac{1}{x^2 - x + 1} dx = \Lambda_1 + \Lambda_2,$$

and we need to evaluate the integrals  $\Lambda_1$  and  $\Lambda_2$ . We have (up to a constant)

$$\Lambda_1 = \frac{1}{2} \int \frac{2x - 1}{x^2 - x + 1} dx \stackrel{t=x^2-x+1}{\underset{dt=(2x-1)dx}{=}} \frac{1}{2} \int \frac{dt}{t} = \frac{1}{2} \ln |t| = \frac{1}{2} \ln(t) = \frac{1}{2} \ln(x^2 - x + 1). \quad (\alpha)$$

Next

$$\begin{aligned} \Lambda_2 &= \frac{7}{2} \int \frac{dx}{x^2 - x + 1} = \frac{7}{2} \int \frac{dx}{\left(x - \frac{1}{2}\right)^2 + \frac{3}{4}} = \frac{7}{2} \int \frac{dx}{\left(x - \frac{1}{2}\right)^2 + \left(\frac{\sqrt{3}}{2}\right)^2} \stackrel{u=x-\frac{1}{2}}{\underset{du=dx}{=}} \frac{7}{2} \int \frac{du}{u^2 + \left(\frac{\sqrt{3}}{2}\right)^2} = \\ &= \frac{7}{2} \cdot \frac{1}{\frac{\sqrt{3}}{2}} \arctan\left(\frac{u}{\frac{\sqrt{3}}{2}}\right) = \frac{7}{\sqrt{3}} \arctan\left(\frac{2u}{\sqrt{3}}\right) = \frac{7\sqrt{3}}{3} \arctan\left(\frac{\sqrt{3}(2x-1)}{3}\right). \quad (\beta) \end{aligned}$$

A combination of  $(\alpha)$  and  $(\beta)$  gives

$$\Lambda = \Lambda_1 + \Lambda_2 = \frac{1}{2} \ln(x^2 - x + 1) + \frac{7\sqrt{3}}{3} \arctan\left(\frac{\sqrt{3}(2x-1)}{3}\right). \quad (A)$$

In the second integral  $M := \int \frac{x - 6}{(x^2 - x + 1)^2} dx$  the dominator appears in the second power. Hence one can apply a similar technique to those presented in 6.B.24. We have

$$M = \int \frac{x - 6}{(x^2 - x + 1)^2} dx = \frac{1}{2} \int \frac{2x - 1}{(x^2 - x + 1)^2} dx - \frac{11}{2} \int \frac{dx}{(x^2 - x + 1)^2} = M_1 + M_2.$$

Then, up to a constant we get

$$M_1 = \frac{1}{2} \int \frac{2x-1}{(x^2-x+1)^2} dx \stackrel{t=x^2-x+1}{dt=(2x-1)dx} \frac{1}{2} \int \frac{dt}{t^2} = -\frac{1}{2t} = -\frac{1}{2(x^2-x+1)}, \quad (\gamma)$$

and

$$\begin{aligned} M_2 &= -\frac{11}{2} \int \frac{dx}{(x^2-x+1)^2} = -\frac{11}{2} \int \frac{dx}{\left(\left(x-\frac{1}{2}\right)^2 + \left(\frac{\sqrt{3}}{2}\right)^2\right)^2} = -\frac{11}{2} \cdot K_2 \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) = \\ &= -\frac{11}{2} \cdot \frac{1}{\frac{3}{4}} \left( \frac{1}{2} K_1 \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) + \frac{x-\frac{1}{2}}{2\left(\left(x-\frac{1}{2}\right)^2 + \frac{3}{4}\right)} \right) = -\frac{11}{2} \cdot \frac{4}{3} \left( \frac{1}{2} \cdot \frac{1}{\frac{\sqrt{3}}{2}} \arctan \left(\frac{x-\frac{1}{2}}{\frac{\sqrt{3}}{2}}\right) + \frac{\frac{2x-1}{2}}{2(x^2-x+1)} \right) = \\ &= -\frac{22}{3} \left( \frac{\sqrt{3}}{3} \arctan \left(\frac{\sqrt{3}(2x-1)}{3}\right) + \frac{2x-1}{4(x^2-x+1)} \right) = -\frac{22\sqrt{3}}{9} \arctan \left(\frac{\sqrt{3}(2x-1)}{3}\right) - \frac{11(2x-1)}{6(x^2-x+1)}. \quad (\delta) \end{aligned}$$

Therefore by  $(\gamma)$  and  $(\delta)$  we obtain

$$M = M_1 + M_2 = -\frac{1}{2(x^2-x+1)} - \frac{22\sqrt{3}}{9} \arctan \left(\frac{\sqrt{3}(2x-1)}{3}\right) - \frac{11(2x-1)}{6(x^2-x+1)}. \quad (B)$$

Combining now  $(\star)$  with  $(A)$  and  $(B)$  we finally deduce that

$$\begin{aligned} \int Q(x) dx &= \ln|x| + \frac{1}{2} \ln(x^2-x+1) + \frac{7\sqrt{3}}{3} \arctan \left(\frac{\sqrt{3}(2x-1)}{3}\right) - \frac{1}{2(x^2-x+1)} \\ &\quad - \frac{22\sqrt{3}}{9} \arctan \left(\frac{\sqrt{3}(2x-1)}{3}\right) - \frac{11(2x-1)}{6(x^2-x+1)} + C, \end{aligned}$$

or equivalently

$$\int Q(x) dx = \ln \left| x \sqrt{x^2-x+1} \right| - \frac{\sqrt{3}}{9} \arctan \left(\frac{\sqrt{3}(2x-1)}{3}\right) - \frac{11x-4}{3(x^2-x+1)} + C,$$

for some constant  $C$ . For a confirmation (and much faster computation) via Sage, just type

```
Q(x)=(2*x^4+2*x^2-5*x+1)/(x*(x^2-x+1)^2); show(Q.integrate(x))
```

Or you may like to verify the integrals  $A_1, A_2, M_1, M_2$  presented above, which can be done in a similar way:

```
show((1/2)*integral((2*x-1)/(x^2-x+1), x)); show((7/2)*integral(1/(x^2-x+1), x))
show((1/2)*integral((2*x-1)/(x^2-x+1)^2, x)); show((-11/2)*integral(1/(x^2-x+1)^2, x))
```

**6.B.28.** An appropriate substitution is given by  $t = e^x$ , hence  $x = \ln(t)$  and  $dx = \frac{1}{t} dt$ . This will allow us to convert the function  $Q(x) = \frac{1}{e^{2x}-4e^x}$  to a rational function, hence we can apply the theory described above. In particular,

$$\int Q(x) dx = \int \frac{1}{e^{3x}-2e^{2x}} dx = \int \frac{dt}{(t^3-2t^2)t} = \int \frac{dt}{t^3(t-2)}.$$

Now, suppose that

$$\frac{1}{t^3(t-2)} = \frac{A}{t} + \frac{B}{t^2} + \frac{C}{t^3} + \frac{D}{t-2}, \quad \text{for some } A, B, C, D \in \mathbb{R}.$$

This gives the relation

$$1 = t^2(t-2)A + t(t-2)B + (t-2)C + t^3D \iff 1 = t^3(A+D) + t^2(-2A+B) + t(-2B+C) - 2C.$$

Thus we get the system

$$-2C = 1, \quad -2B + C = 0, \quad -2A + B = 0, \quad A + D = 0,$$

which has the tuple  $\{A = -1/8, B = -1/4, C = -1/2, D = 1/8\}$  as a unique solution. This means that

$$\frac{1}{t^3(t-2)} = -\frac{1}{8t} - \frac{1}{4t^2} - \frac{1}{2t^3} + \frac{1}{8(t-2)}, \quad (*)$$

which can be easily confirmed in Sage by the usual method, that is,

```
var("t"); show((1/(t^4-2*t^3)).partial_fraction())
```

Based on  $(*)$  we may finally compute the integral at hand:

$$\int \frac{dt}{t^3(t-2)} = \int \left( \frac{1}{8(t-2)} - \frac{1}{8t} - \frac{1}{4t^2} - \frac{1}{2t^3} \right) dt = \frac{1}{8} (\ln|t-2| - \ln|t|) + \frac{1}{4t} + \frac{1}{4t^2} + C =$$

$$= \frac{1}{8} (\ln |e^x - 2| - \ln(e^x)) + \frac{1}{4e^x} \left(1 + \frac{1}{e^x}\right) + C = \frac{\ln |e^x - 2|}{8} - \frac{x}{8} + \frac{1}{4} e^{-2x} (1 + e^x) + C,$$

for some constant  $C$ . For a quick confirmation of this result try the command

```
show(integral(1/(e**(3*x)-2*e**(2*x)), x))
```

**6.B.31.** For all  $x \neq \frac{\pi}{2} + k\pi$ , with  $k \in \mathbb{Z}$ , we see that

$$\int \frac{1}{\cos^2(x)} dx = \int_{dt=dx/\cos^2(x)}^{t=\tan(x)} dt = t + C = \tan(x) + C,$$

for some constant  $C$ . Therefore, with the help of the table given in 5.A.3 one computes

$$\begin{aligned} \int_{\pi/6}^{\pi/3} \tan^2(x) dx &= \int_{\pi/6}^{\pi/3} \frac{\sin^2(x)}{\cos^2(x)} dx = \int_{\pi/6}^{\pi/3} \frac{1 - \cos^2(x)}{\cos^2(x)} dx = \int_{\pi/6}^{\pi/3} \left( \frac{1}{\cos^2(x)} - 1 \right) dx = \left[ \tan(x) - x \right]_{\pi/6}^{\pi/3} \\ &= \sqrt{3} - \frac{\pi}{3} - \left( \frac{1}{\sqrt{3}} - \frac{\pi}{6} \right) = \frac{2\sqrt{3}}{3} - \frac{\pi}{6}. \end{aligned}$$

An alternative for computing this integral occurs by the substitution  $u = \tan(x)$ , with

$$du = \frac{dx}{\cos^2(x)}, \quad \sin^2(x) = \frac{\tan^2(x)}{1 + \tan^2(x)} = \frac{u^2}{1 + u^2}.$$

This is also based on the integral  $\int \frac{1}{1+u^2} du = \arctan(u) + C$  and left as an easy challenge.

Notice, to confirm the result in Sage we may type

```
show(integral(tan(x)*tan(x), x, pi/6, pi/3))
```

For the second integral recall by 6.B.8 that integration by parts gives

$$\int \frac{x}{\cos^2(x)} dx = x \tan(x) + \ln |\cos(x)| + C,$$

for all  $x \neq \frac{\pi}{2} + k\pi$ ,  $k \in \mathbb{Z}$ , where  $C$  is some constant. Thus

$$\int_0^{\pi/4} \frac{x}{\cos^2(x)} dx = \left[ x \tan(x) + \ln(\cos(x)) \right]_0^{\pi/4} = \frac{\pi}{4} + \ln \frac{\sqrt{2}}{2} = \frac{\pi}{4} - \frac{\ln(2)}{2}.$$

Again we can confirm this result in Sage, as before, i.e.,

```
show(integral(x/(cos(x)*cos(x)), x, 0, pi/4))
```

**6.B.32.** (a) Set  $y = 1 - x^2$  with  $dy = -2dx$ . For  $x = 0$  we have  $y = 1$  and for  $x = 1$  we have  $y = 0$ . Thus

$$\int_0^1 \frac{x}{\sqrt{1-x^2}} dx = - \int_1^0 \frac{y^{-1/2}}{2} dy = \int_0^1 \frac{y^{-1/2}}{2} dy = \left[ \sqrt{y} \right]_0^1 = 1.$$

(b) Set  $t = x + \sqrt{x^2 - 1}$ , so that  $dt = \frac{\sqrt{x^2-1}+x}{\sqrt{x^2-1}} dx$ . For  $x = 1$  we get  $t = 1$  and for  $x = 2$  we get  $t = 2 + \sqrt{3}$ . Thus,

$$\int_1^2 \frac{dx}{\sqrt{x^2-1}} = \int_1^{2+\sqrt{3}} \frac{1}{t} dt = \left[ \ln(t) \right]_1^{2+\sqrt{3}} = \ln(2 + \sqrt{3}).$$

**6.B.33.** Using Sage we see that the integral at hand equals to  $-1/2$ . This can be done in the usual way, i.e., the cell

```
integral(sin(3*x)*cos(x), x, pi/4, pi)
```

One method to verify this result formally is based on the identity  $\sin(\alpha + \beta) + \sin(\alpha - \beta) = 2 \sin(\alpha) \cos(\beta)$  which for our case applies as  $2 \sin(3x) \cos(x) = \sin(3x + x) + \sin(3x - x)$ . Thus we get

$$\begin{aligned} \int_{\pi/4}^{\pi} \sin(3x) \cos(x) dx &= \frac{1}{2} \int_{\pi/4}^{\pi} (\sin(3x + x) + \sin(3x - x)) dx = \frac{1}{2} \int_{\pi/4}^{\pi} \sin(4x) dx + \frac{1}{2} \int_{\pi/4}^{\pi} \sin(2x) dx \\ &= -\frac{1}{8} \left[ \cos(4x) \right]_{\pi/4}^{\pi} - \frac{1}{4} \left[ \cos(2x) \right]_{\pi/4}^{\pi} = -\frac{1}{8} (1 - (-1)) - \frac{1}{4} (1 - 0) = -\frac{1}{4} - \frac{1}{4} = -\frac{1}{2}. \end{aligned}$$

**6.B.34.** (a) Because  $0 \leq \frac{x^9}{\sqrt{2}} \leq \frac{x^9}{\sqrt{1+x}} \leq x^9$  for all  $x \in [0, 1]$  the geometric meaning of the definite integral implies

$$\frac{\sqrt{2}}{20} = \int_0^1 \frac{x^9}{\sqrt{2}} dx \leq \int_0^1 \frac{x^9}{\sqrt{1+x}} dx \leq \int_0^1 x^9 dx = \frac{1}{10}.$$

(b) Because  $0 < x < \pi/2$  we have  $0 < \frac{2}{\pi} < \frac{1}{x}$ . Also we have  $0 \leq \sin(x) \leq 1$  for all  $x \in (0, \pi/2)$ . Thus

$$\int_0^{\pi/2} \frac{2}{\pi} dx < \int_0^{\pi/2} \frac{\sin(x)}{x} < \int_0^{\pi/2} 1 dx,$$

or equivalently  $\frac{2}{\pi} \cdot \frac{\pi}{2} < \int_0^{\pi/2} \frac{\sin(x)}{x} < \frac{\pi}{2}$ .

**6.B.35.** Recall that any continuous function  $f : [-a, a] \rightarrow \mathbb{R}$  satisfies the relation

$$\int_{-a}^a f(x) dx = \int_{-a}^0 f(x) dx + \int_0^a f(x) dx. \quad (*)$$

Set  $u = -x$  such that  $du = -dx$ . For  $x = -a$  we have  $u = a$  and for  $x = 0$  we have  $u = 0$ . Assume now that  $f$  is even. Then  $f(u) = f(-x) = f(x)$  and thus

$$\int_{-a}^0 f(x) dx = \int_a^0 f(u)(-du) = - \int_a^0 f(u) du = \int_0^a f(u) du = \int_0^a f(x) dx$$

and hence  $(*)$  gives  $\int_{-a}^a f(x) dx = 2 \int_0^a f(x) dx$ . Similarly is treated the case where  $f$  is odd.

**6.B.36.** (1) Since  $f$  is periodic with period  $T$ , for any  $a \in \mathbb{R}$  one can write

$$\int_a^{a+T} f(x) dx = \int_a^0 f(x) dx + \int_0^T f(x) dx + \int_T^{a+T} f(x-T) dx. \quad (\dagger)$$

Set now  $u = x - T$  with  $du = dx$ . For  $x = T$  we have  $u = 0$  and for  $x = a + T$  we have  $u = a$ . Thus,  $\int_T^{a+T} f(x-T) dx = \int_0^a f(u) du$ , and the result follows by  $(\dagger)$ .

An alternative relies on the primitive function  $F$  of  $f$ , which allows us to prove that

$$\begin{aligned} \int_0^T f(x) dx &= \int_0^a f(x) dx + \int_a^T f(x) dx = \int_0^a f(x) dx + F(T) - F(a) \\ &= \int_0^a f(x) dx + (F(a+T) - F(a)) - (F(a+T) - F(T)) \\ &= \int_0^a f(x) dx + \int_a^{a+T} f(x) dx - \int_T^{a+T} f(x) dx. \quad (\ddagger) \end{aligned}$$

Then, as above with the substitution  $t = x - T$  one can show that  $\int_T^{a+T} f(x) dx = \int_0^a f(x) dx$  and the result follows by  $(\ddagger)$ .

(2) Since  $f$  is periodic with period  $T$ , using induction over the naturals we can show that  $f(x) = f(x + nT) = f(x - nT)$ , for any  $n \in \mathbb{N}$ . This easily extends to  $n \in \mathbb{Z}$  (try to prove this claim). Using this periodicity property, for  $n \geq 0$  we get

$$\begin{aligned} \int_a^{a+nT} f(x) dx &= \int_a^0 f(x) dx + \sum_{k=0}^{n-1} \int_{kT}^{(k+1)T} f(x) dx + \int_{nT}^{a+nT} f(x) dx \\ &= \int_a^0 f(x) dx + \sum_{k=0}^{n-1} \int_{kT}^{(k+1)T} f(x - kT) dx + \int_{nT}^{a+nT} f(x - nT) dx \\ &= \int_a^0 f(x) dx + \sum_{k=0}^{n-1} \int_0^T f(x) dx + \int_0^T f(x) dx = \sum_{k=0}^{n-1} \int_0^T f(x) dx = n \int_0^T f(x) dx. \quad (*) \end{aligned}$$

Notice for the integral  $\int_{kT}^{(k+1)T} f(x - kT) dx$  we did the substitution  $t = x - kT$  with  $dt = dx$ , and  $t = 0$  for  $x = kT$ ,

$t = T$  for  $x = (k + 1)T$ . Similarly for  $\int_{nT}^{a+nT} f(x - nT) dx$ .

Suppose now that  $n < 0$ . Then we have

$$\int_a^{a+nT} f(x) dx = - \int_{a+nT}^a f(x) dx = - \int_{a+nT}^{a+nT+(-nT)} f(x) dx \stackrel{(*)}{=} - \left( -n \int_0^T f(x) dx \right) = n \int_0^T f(x) dx.$$

This proves our claim. Observe, that the identity in (1) for  $a = 0$  reduces to

$$\int_0^{nT} f(x) dx = n \int_0^T f(x) dx.$$

This says that the area under  $C_f$  for  $n$  periods equals  $n$  times the area under  $C_f$  for one period.

**6.B.38.** (a) The given function  $F$  is clearly the antiderivative of the function  $f(x) := x^5 \ln(x + 1)$  on the interval  $(-1, 1)$ . Thus,  $F'(x) = -x^5 \ln(x + 1)$ , see also the fundamental theorem of calculus in 6.2.9.

(b) We have  $F'(x) = x^2 \sin(x) + 4 \cos(4x)$  and we see that

$$|g(x)| = \left| \frac{x^2 \sin(x) + 4 \cos(4x)}{x^2 + 2} \right| \leq \frac{|x^2 \sin(x)| + 4|\cos(4x)|}{x^2 + 2} \leq \frac{|x^2| \cdot 1 + 4 \cdot 1}{x^2 + 2} = \frac{x^2 + 4}{x^2 + 1},$$

for all  $x \in (0, +\infty)$ . Thus, obviously  $\lim_{x \rightarrow \infty} \frac{x^2+4}{x^2+1} = 1$ , which you can confirm in Sage by the cell

```
var("x"); lim((x^2+4)/(x^2+1), x=oo)
```

**6.B.45.** For some real numbers  $a < b$  consider the set  $A = (a, b]$  and the function  $f : A \rightarrow \mathbb{R}$ , with  $f(x) = 1/(x - a)$ . Though  $f$  is continuous on  $A$ , it is not uniformly continuous, since  $A$  is *not* closed. Indeed, let  $(x_n)$  and  $(y_n)$  be the sequences defined by

$$x_n = a + \frac{b - a}{n}, \quad y_n = a + \frac{b - a}{n + 1}, \quad n \in \mathbb{Z}_+.$$

Then it is easy to see that  $(x_n - y_n) \rightarrow 0$  but  $f(x_n) - f(y_n)$  does not converge to 0. Hence, by 6.B.44 one deduces that  $f$  cannot be uniformly continuous on  $A$ .

As an example where the domain is *not* bounded, consider the parabola  $h : A \rightarrow \mathbb{R}$ ,  $h(x) = x^2$  defined on  $A = [0, +\infty)$ . The sequences  $(x_n = n)$  and  $(y_n = n - \frac{1}{n})$  with  $n \in \mathbb{Z}_+$ , they both belong to  $A$  and satisfy  $x_n - y_n = \frac{1}{n} \rightarrow 0$ , as  $n \rightarrow \infty$ . However,  $h(x_n) - h(y_n) = 2 - \frac{1}{n^2}$ , which obviously does not converge to 0. By 6.B.44 we deduce that  $h$  cannot be uniformly continuous on  $A$ .

**6.B.46.** Consider the sequences  $(x_n), (y_n)$  with general terms  $x_n = 1/(n + 1)$  and  $y_n = 1/n$ , respectively, with  $n \geq 2$ . They both belong to  $A = (0, 1)$  and satisfy  $x_n - y_n \rightarrow 0$ . However, it is easy to see that  $f(x_n) - f(y_n) = 1$  and hence  $\lim_{n \rightarrow \infty} (f(x_n) - f(y_n)) = 1 \neq 0$ . The assertion now follows by the statement in 6.B.44. Provide an alternative proof based on the  $\varepsilon$ - $\delta$ -definition of uniform continuity, given in 6.2.11.

**6.B.48.** The proof is analogous to 6.B.47 and we leave it to the reader for practice.

**6.B.49.** (a) For instance, consider the function  $f(x) = \sin\left(\frac{1}{x}\right)$  with  $x \in A = (0, 1]$ . It is not hard to see that  $f$  is continuous on  $A$ , but not uniformly continuous. Now, for any  $n \in \mathbb{N}$  we have  $n\pi + \frac{\pi}{2} \geq 1$  and hence the sequence  $(x_n)$  with general term

$$x_n = \frac{1}{n\pi + \pi/2}, \quad n \in \mathbb{N},$$

is a sequence in  $A$ . Since  $n\pi + \pi/2 \rightarrow +\infty$  we have  $x_n \rightarrow 0$ , as  $n \rightarrow +\infty$ . Thus  $(x_n)$  is convergent and thus a Cauchy sequence in  $A$ . However, the sequence  $(f(x_n)) = \left(\sin\left(\frac{1}{x_n}\right)\right)_{n \in \mathbb{N}}$  is not a Cauchy sequence. Why?.

(b) Let  $(x_n)$  be a Cauchy sequence with  $x_n \in A$  for all  $n$ . By assumption  $f$  is uniformly continuous on  $A$ , hence for every  $\varepsilon > 0$  there exists  $\delta = \delta(\varepsilon) > 0$  such that  $|f(x) - f(y)| < \varepsilon$ , for all  $x, y \in A$ , provided that  $|x - y| < \delta$ . However,  $(x_n)$  is Cauchy and since  $\delta > 0$  we can find some natural  $N = N(\delta)$  such that  $|x_n - x_m| < \delta$  for all  $n, m \geq N$ . But then we get  $|f(x_n) - f(x_m)| < \varepsilon$ , for all  $n, m \geq N$ , and the result follows.

**6.B.50.** The function  $f(x) = x^2$  is continuous on  $\mathbb{R}$  and hence also on  $A = [0, 1]$ . Since the set  $[0, 1]$  is a closed finite interval, the function  $f$  should be uniformly continuous, see the theorem in 6.2.11.

Suppose that the function  $g(x) = \tan(x)$  is uniformly continuous on  $B = [0, \pi/2)$ . The domain of  $g$  is a finite subset of  $\mathbb{R}$  and hence bounded. But then the function  $g$  should be also bounded.<sup>6</sup> However, it is easy to see that  $\tan(x_n) \rightarrow +\infty$  for  $x_n \in (0, \pi/2)$  with  $x_n \rightarrow \pi/2$ , which implies that  $g(x) = \tan(x)$  is not bounded (think also the graph of  $\tan(x)$ ). This gives a contradiction and hence  $g(x) = \tan(x)$  is not uniformly continuous on  $B$ .

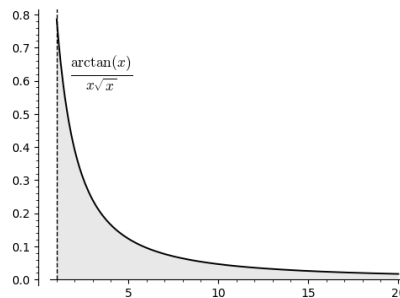
Next, the function  $h(x) = x^2$  is not uniformly continuous on  $\mathbb{R}$ . For instance, consider the sequences  $(x_n = n)$  and  $(y_n = n + \frac{1}{n})$ , with  $x_n - y_n \rightarrow 0$ . Then we see that  $h(x_n) - h(y_n) = 2 + \frac{1}{n^2} \geq 2$  and thus  $\lim_{n \rightarrow +\infty} (h(x_n) - h(y_n)) \neq 0$ . Our assertion now follows by 6.B.44.

Finally, the function  $k(x) = x^3$  is also not uniformly continuous on  $\mathbb{R}$ . Here one may consider the sequences  $x_n = n + \frac{1}{n}$  and  $y_n = n$ . Then obviously  $x_n - y_n \rightarrow 0$  but we see that  $k(x_n) - k(y_n) \geq 3n$  for all  $n$ , which implies that  $\lim_{n \rightarrow +\infty} (k(x_n) - k(y_n)) = +\infty$ .

**6.B.52.** The improper integral represents the area of the region bounded by the graph of the positive function

$$f(x) = \frac{\arctan(x)}{x\sqrt{x}} = x^{-\frac{3}{2}} \arctan(x), \quad x \geq 1,$$

and the  $x$ -axis. From the left, this region is bounded by the line  $x = 1$ , see also the figure here:



Therefore, the integral is a positive real number, or equals  $+\infty$ . First we see that

$$\int_1^{+\infty} x^{-\frac{3}{2}} dx = \lim_{t \rightarrow +\infty} \int_1^t x^{-\frac{3}{2}} dx = \lim_{t \rightarrow +\infty} \left(2 - \frac{2}{\sqrt{t}}\right) = 2.$$

Moreover, we know that

$$\frac{\pi}{4} \leq \arctan(x) \leq \frac{\pi}{2}, \quad \forall x \in [1, +\infty),$$

and hence we get

$$\frac{\pi}{2} = \frac{\pi}{4} \int_1^{+\infty} x^{-\frac{3}{2}} dx \leq \int_1^{+\infty} \frac{\arctg x}{x\sqrt{x}} dx \leq \frac{\pi}{2} \int_1^{+\infty} x^{-\frac{3}{2}} dx = \pi.$$

Thus the integral at hand is a finite real number. In fact, with the aid of Sage one can compute this integral via the command `integral(arctan(x)/(x*sqrt(x)), x, 1, oo)`.

**6.B.57.** As usual you can use the `def` method and introduce a routine which you may call “`length_of_curve`”. Since we usually work with the variable  $t$  for a parametric curve  $c : [a, b] \rightarrow \mathbb{R}^2$ ,  $t$  should be declared as a symbolic variable, together with the endpoints  $a, b$ , that is, the limits of integration. Thus, the input of our routine can be a tuple  $(x, y, a, b)$ , corresponding to the curve  $c : [a, b] \rightarrow \mathbb{R}^2$  with  $c(t) = [x(t), y(t)]$  for all  $t \in [a, b]$ . Hence, though  $a, b$  are real number inputs,  $x, y : [a, b] \rightarrow \mathbb{R}$  are real-valued functions of  $t$  which it is necessary to introduce as symbolic functions. To control also the aspect ratio of the corresponding parametric plot (and take a better illustration), we will introduce one more symbolic variable, say `rt`, which for any curve we can specify so that the result will be the desired one. This means that finally the input of our routine will be the list  $(x, y, a, b, rt)$ , and the implementation goes as follows:

```
var("t, a, b, rt")
function("x")(t)
function("y")(t)
def length_of_curve(x, y, a, b, rt) :
    X(t)=diff(x, t)
```

<sup>6</sup>Show that any uniformly-continuous function  $f : A \rightarrow \mathbb{R}$  mapping a bounded interval to  $\mathbb{R}$ , is bounded.

```

Y(t)=diff(y, t)
l(t)=(X(t)*X(t)+Y(t)*Y(t)).factor()
s(t)=numerical_approx(integral(sqrt(l), t, a, b), digits=7)
show("The length of the curve c(t)=(x, y) on [a, b] equals to ", ":", s(t))
p=parametric_plot((x, y), (t, a, b))
p.show(figsize=6, aspect_ratio=rt)
return
    
```

To test the routine first we will confirm a known result based on the curve  $[e^t - t, 4e^{t/2}]$  of the previous task (see 6.B.56), with  $t \in [0, 1]$ . To confirm this case, now one can simply give the cell

```
length_of_curve(e^t-t, 4*e^(t/2), 0, 1, 1/2)
```

where notice that we fixed  $rt = 1/2$ . Sage returns the following answer:

The length of the curve  $c(t) = (-t + e^t, 4e^{(\frac{1}{2}t)})$  on  $[0, 1]$  equals to:2.718282

For the given cases one can proceed similarly (we present them case by case):

(1) For the curve  $c(t) = [\cos^3(t), \sin^3(t)]$  on  $[0, \pi/2]$  type

```
length_of_curve((cos(t))^3, (sin(t))^3, 0, pi/2, 1/2)
```

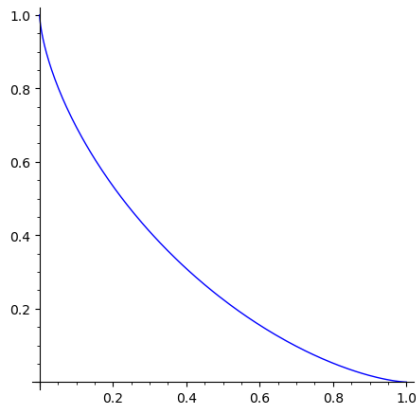
Here the output is

The length of the curve  $c(t) = (\cos(t)^3, \sin(t)^3)$  on  $[0, \frac{1}{2}\pi]$  equals to:1.500000

Indeed, we see that  $|c'(t)| = \sqrt{9\cos^4(t)\sin^2(t) + 9\sin^4(t)\cos^2(t)} = 3\cos(t)\sin(t)$  and thus

$$s_{c(t)} = \int_0^{\pi/2} 3\cos(t)\sin(t) dt = \frac{3}{2} \int_0^{\pi/2} \sin(2t) dt = \frac{3}{2} \left[ -\frac{\cos(2t)}{2} \right]_0^{\pi/2} = \frac{3}{2} \left[ -\frac{\cos(\pi)}{2} + \frac{\cos(0)}{2} \right] = \frac{3}{2}.$$

As for the graph of  $c(t)$ , Sage returns the figure given here:



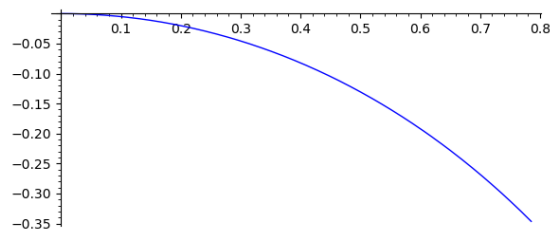
(2) For the curve  $c(t) = [t, \ln(\cos(t))]$  on  $[0, \pi/4]$  type

```
length_of_curve(t, ln(cos(t)), 0, pi/4, 1)
```

Sage's output has the form

The length of the curve  $c(t) = (t, \log(\cos(t)))$  on  $[0, \frac{1}{4}\pi]$  equals to:0.8813736

and the corresponding plot has the form



To confirm Sage's result we compute

$$|c'(t)| = \sqrt{1 + \left(\frac{-\sin(t)}{\cos(t)}\right)^2} = \sqrt{1 + \tan^2(t)} = \sqrt{\sec^2(t)} = \sec(t),$$

where we used the identity  $\tan^2(x) = \sec^2(x) - 1$ . Thus

$$s_{c(t)} = \int_0^{\pi/4} |c'(t)| dt = \int_0^{\pi/4} \sec(t) dt.$$

Notice now that

$$\int \sec(x) dx = \int \frac{\sec(x)(\sec(x) + \tan(x))}{\sec(x) + \tan(x)} dx = \int \frac{\sec^2(x) + \sec(x)\tan(x)}{\sec(x) + \tan(x)} dx.$$

Setting  $u = \sec(x) + \tan(x)$  we compute

$$\begin{aligned} du &= (\sec(x) + \tan(x))' dx = \left( \left(\frac{1}{\cos(x)}\right)' + \tan'(x) \right) dx = \left( \frac{\sin(x)}{\cos^2(x)} + (1 + \tan^2(x)) \right) dx = \\ &= \left( \frac{1}{\cos(x)} \cdot \frac{\sin(x)}{\cos(x)} + \sec^2(x) \right) dx = (\sec(x)\tan(x) + \sec^2(x)) dx, \end{aligned}$$

and hence

$$\int \sec(x) dx = \int \frac{1}{u} du = \ln |u| + C, \quad C \in \mathbb{R}. \quad (\star)$$

Since we have  $u(0) = \sec(0) + \tan(0) = 1$  and  $u(\pi/4) = \sec(\pi/4) + \tan(\pi/4) = \sqrt{2} + 1$ , we finally deduce that

$$s_{c(t)} = \int_0^{\pi/4} \sec(t) dt = \int_1^{\sqrt{2}+1} \frac{du}{u} = \left[ \ln |u| \right]_1^{\sqrt{2}+1} = \ln(1 + \sqrt{2}) - \ln(1) = \ln(1 + \sqrt{2}) \approx 0.8813736.$$

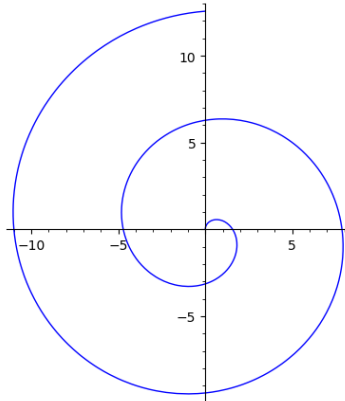
(3) For the curve  $c(t) = [t \sin(t), t \cos(t)]$  on  $[0, 4\pi]$  type

```
length_of_curve(t*sin(t), t*cos(t), 0, 4*pi, 1)
```

which answers

The length of the curve  $c(t) = (t \sin(t), t \cos(t))$  on  $[0, 4\pi]$  equals to: 80.81931

and prints out the following figure



Let us present some hints helpful for verifying this result formally. First we compute

$$|c'(t)| = \sqrt{(\sin(t) + t \cos(t))^2 + (\cos(t) - t \sin(t))^2} = \sqrt{1 + t^2}$$

and thus we need to compute  $s_{c(t)} = \int_0^{4\pi} \sqrt{1 + t^2} dt$ . For this integral, using the relation  $(\star)$  from (2) we will first prove that

$$\int \sqrt{1 + x^2} dx = \frac{1}{2} x \sqrt{1 + x^2} + \frac{1}{2} \ln(x + \sqrt{1 + x^2}) + C. \quad (*)$$

Indeed, set  $x = \tan(\theta)$  with  $dx = \left(\frac{1}{\cos^2(x)}\right) d\theta = \sec^2(x) d\theta$ . Then

$$\int \sqrt{1 + x^2} dx = \int \sqrt{1 + \tan^2(\theta)} \sec^2(\theta) d\theta = \int \sqrt{\sec^2(\theta)} \sec^2(\theta) d\theta = \int \sec^3(\theta) d\theta.$$



We have  $(\tan(\theta))' = \sec^2(\theta)$  and  $(\sec(\theta))' = \sec(\theta) \tan(\theta)$  and integration by parts gives

$$\begin{aligned} \int \sec^3(\theta) d\theta &= \int \sec^2(\theta) \sec(\theta) d\theta = \int (\tan(\theta))' \sec(\theta) d\theta = \tan(\theta) \sec(\theta) - \int \tan(\theta) (\sec(\theta))' d\theta \\ &= \tan(\theta) \sec(\theta) - \int \tan^2(\theta) \sec(\theta) d\theta = \tan(\theta) \sec(\theta) - \int (\sec^2(\theta) - 1) \sec(\theta) d\theta \\ &= \tan(\theta) \sec(\theta) - \int \sec^3(\theta) d\theta + \int \sec(\theta) d\theta. \end{aligned}$$

Thus  $\int \sec^3(\theta) d\theta = \frac{1}{2}(\tan(\theta) \sec(\theta) + \int \sec(\theta) d\theta)$  and hence using (\*) we can write

$$\int \sqrt{1+x^2} dx = \int \sec^3(\theta) d\theta = \frac{1}{2} \tan(\theta) \sec(\theta) + \frac{1}{2} \ln |\sec(\theta) + \tan(\theta)| + C.$$

The desired formula (\*) appears now by replacing  $x = \tan(\theta)$  and  $\sec(\theta) = \sqrt{1+x^2}$ . The final computation of  $s_{c(t)}$  is now based on (\*) and left as an easy exercise.

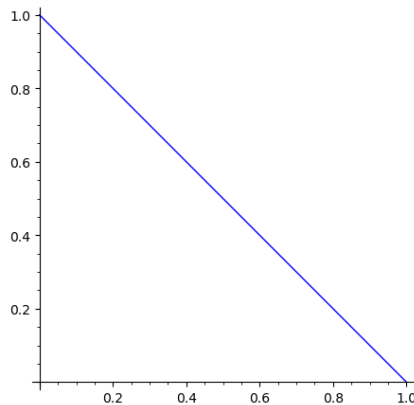
(4) For the curve  $c(t) = [\sin^2(t), \cos^2(t)]$  with  $t \in [0, \frac{\pi}{2}]$ , similarly we the previous cases we can type

```
length_of_curve((sin(t))^2, (cos(t))^2, 0, pi/2, 1)
```

Sage returns<sup>7</sup>

The length of the curve  $c(t) = (\sin^2(t), \cos^2(t))$  on  $[0, \frac{1}{2}\pi]$  equals to: 1.414214

and the following illustration of the curve at hand:



For a formal computation, we see that

$$|c'(t)| = \sqrt{(2 \sin(t) \cos(t))^2 + (-2 \sin(t) \cos(t))^2} = \sqrt{(\sin(2t))^2 + (-\sin(2t))^2} = \sqrt{2 \sin^2(2t)}.$$

Thus

$$s_{c(t)} = \int_0^{\pi/2} |c'(t)| dt = \sqrt{2} \int_0^{\pi/2} \sin(2t) dt = \sqrt{2} \left[ -\frac{\cos(2t)}{2} \right]_0^{\pi/2} = \sqrt{2}.$$

In fact, the given curve is a part of the line  $y = 1 - x$  (since  $\sin^2(t) + \cos^2(t) = 1$ ), and in particular the segment with boundary points  $[0, 1]$  for  $t = 0$ , and  $[1, 0]$  for  $t = \frac{\pi}{2}$ , see also the figure given above. Hence one can immediately write its length,  $s_{c(t)} = \sqrt{2}$  (by the Pythagorean theorem).

**6.B.58.** It is easy to see that

$$f'(t) = -r \sin t + \frac{r}{2 \tan \frac{t}{2} \cdot \cos^2 \frac{t}{2}} = -r \sin t + \frac{r}{\sin t} = \frac{r \cos^2 t}{\sin t}, \quad g'(t) = r \cos t,$$

for any  $t \in [\pi/2, a]$ . Thus, for the length  $s_{\alpha(t)}$  we get

$$s_{\alpha(t)} = \int_{\pi/2}^a \sqrt{\frac{r^2 \cos^4 t}{\sin^2 t} + r^2 \cos^2 t} dt = \int_{\pi/2}^a \sqrt{\frac{r^2 \cos^2 t}{\sin^2 t}} dt = -r \int_{\pi/2}^a \frac{\cos t}{\sin t} dt = -r [\ln(\sin t)]_{\pi/2}^a = -r \ln(\sin a).$$

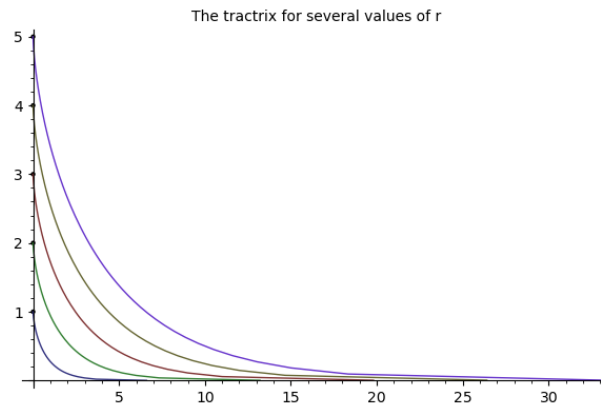
To plot the tractrix for the values  $r = 1, 2, \dots, 5$  one can use the block

<sup>7</sup>Be aware that in this case Sage does not type the curve  $c(t)$  properly.

```

var("r, t")
f(t, r)=r*cos(t)+r*ln(tan(t/2))
g(t, r)=r*sin(t)
a=parametric_plot((f(t, 1), g(t, 1)), (t, pi/2, pi/2+1), exclude=[pi], rgbcolor=(0.2,0.2,0.5),
title="The tractrix for several values of r")
a+=parametric_plot((f(t, 2), g(t, 2)), (t, pi/2, pi/2+1), exclude=[pi], rgbcolor=(0.2,0.5,0.2))
a+=parametric_plot((f(t, 3), g(t, 3)), (t, pi/2, pi/2+1), exclude=[pi], rgbcolor=(0.5,0.2,0.2))
a+=parametric_plot((f(t, 4), g(t, 4)), (t, pi/2, pi/2+1), exclude=[pi], rgbcolor=(0.4,0.4,0.2))
a+=parametric_plot((f(t, 5), g(t, 5)), (t, pi/2, pi/2+1), exclude=[pi], rgbcolor=(0.4,0.2,0.8))
a+=point([[0, r] for r in [1,2,..,5]], color="black")
a.show(aspect_ratio=4, figsize=6)
    
```

which produces the following figure



**6.B.64.** By separating the variables we can write

$$y^{-1/2} dy = e^{\sin(x)} \cos(x) dx,$$

and this implies that

$$\int y^{-1/2} dy = \int e^{\sin(x)} \cos(x) dx \iff 2y^{1/2} = \int e^{\sin(x)} \cos(x) dx + C$$

To compute the integral at the right hand side set  $t = \sin(x)$  with  $dt = \cos(x) dx$ . Then we get

$$2y^{1/2} = \int e^t dt + C \iff 2y^{1/2} = e^t + C = e^{\sin(x)} + C.$$

To compute  $C$  we use the initial condition. By assumption  $y(0) = 16$ , thus  $2 \cdot \sqrt{y(0)} = 2 \cdot \sqrt{16} = 2 \cdot 4 = 8 = e^{\sin(0)} + C$ , from where we get  $C = 7$ . Thus

$$y = y(x) = \left( \frac{e^{\sin(x)} + 7}{2} \right)^2.$$

**6.B.67.** Let us refer to our routine as `trapezoid_rule`. The input should be a function  $f$ , the reals numbers  $a, b$  representing the lower and upper limit of the integral and the number of steps  $n$ . Let us present the program and provide the necessary explanations below.

```

def trapezoid_rule(f, a, b, n):
    step = (b - a) / n
    valsf = [f(a + i * step) for i in range(n + 1)]
    return (step / 2) * (valsf[0] + 2 * sum(valsf[1:n]) + valsf[-1])
    
```

In this block we use the list `valsf`, which is generated by evaluating the function  $f$  at  $n + 1$  points (the latter should be evenly spaced between  $a$  and  $b$ ). Hence, for example, `valsf[i]` contains the function value at the  $i$ th point, i.e., `valsf[i] = f(a + i * step)`. In the return, `valsf[1:n]` selects the sublist containing all the intermediate function values, excluding the first and last points. This is because the trapezoid rule assigns a weight of 2 to the intermediate points between the start and end of the interval, while the first and last points are treated differently with a weight of 1. In our program, these are accessed separately as `valsf[0]` and `valsf[-1]`, respectively.

Let us now test our routine, by verifying, for instance, the result obtained in 6.B.66. It suffices to add the cell

```
n=4; a=0; b=pi; f(x)=sin(x)
print(trapezoid_rule(f, a, b, n).n())
```

Sage's result is 1.89611889793704.

For the estimation of the integral  $L$  by the trapezoid rule, we will first present a formal computation. We have  $a = \pi/2$ ,  $b = 3\pi/2$ ,  $n = 4$  and hence  $h = \pi/4$ . Moreover,

$$x_0 = a = \frac{\pi}{2}, \quad x_1 = x_0 + h = \frac{3\pi}{4}, \quad x_2 = x_0 + 2h = \pi, \quad x_3 = x_0 + 3h = \frac{5\pi}{4}, \quad x_4 = b = x_0 + 4h = \frac{3\pi}{2}.$$

Thus we obtain

$$\begin{aligned} L_{\text{trap}} &= \frac{h}{2} [f(x_0) + 2f(x_1) + 2f(x_2) + 2f(x_3) + f(x_4)] \\ &= \frac{\pi}{8} \left[ \cos\left(\frac{\pi}{2}\right) + 2\cos\left(\frac{3\pi}{4}\right) + 2\cos(\pi) + 2\cos\left(\frac{5\pi}{4}\right) + \cos\left(\frac{3\pi}{2}\right) \right] = \frac{\pi}{8} \left[ 0 - 2\frac{\sqrt{2}}{2} - 2 - 2\frac{\sqrt{2}}{2} + 0 \right] \\ &= -\frac{\pi(1 + \sqrt{2})}{4} \approx -1.896. \end{aligned}$$

We can now confirm this result by applying the `trapezoid_rule` routine:

```
n=4; a=pi/2; b=3*pi/2; f(x)=cos(x)
print(trapezoid_rule(f, a, b, n).n())
```

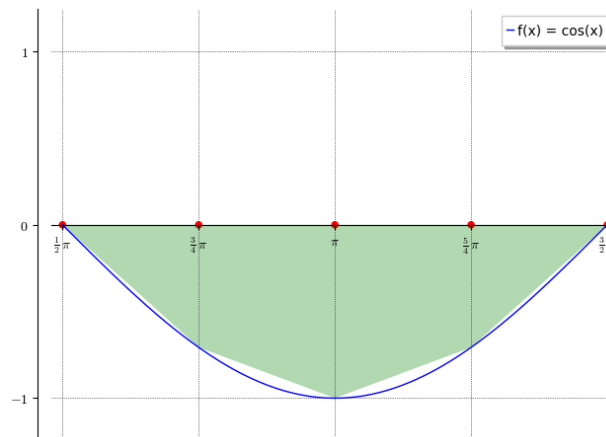
Sage's output is  $-1.89611889793704$

To compute the actual error we first see that

$$L = \int_{\pi/2}^{3\pi/2} \cos(x) dx = [\sin(x)]_{\pi/2}^{3\pi/2} = \sin(3\pi/2) - \sin(\pi/2) = -1 - 1 = -2.$$

Thus,  $|L - L_{\text{trap}}| = |-2 + 1.896| = |-0.104| = 0.104$ . For the theoretical error a simple computation gives that  $\frac{(b-a)h^2}{12} |f''| = \frac{\pi^3}{192} \approx 0.161$ .

It is easy to see that the cosine function is convex over the interval of integration and the trapezoid rule overestimates the integral (since  $-1.896 > -2$ ). This overestimation is characteristic of the trapezoid rule when applied to convex functions. In fact,  $\cos(x)$  is negative on  $[\pi/2, 3\pi/2]$ , and hence the trapezoids do not exceed the area under the curve. This gives the desired overestimation, as illustrated in the figure below.



**6.B.68.** Simpson's rule formula for an integral over  $[a, b]$  with  $n$  intervals, is given by (see 6.2.23)

$$I_{\text{Simp}} = \frac{h}{3} [f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)],$$

where  $h = \frac{b-a}{n}$ ,  $x_i = a + ih$  for  $i = 0, 1, \dots, n$ , and  $n$  is even! Hence, except of the starting and ending points  $x_0, x_n$ , for all odd indices  $i = 1, 3, 5, \dots$  the function values  $f(x_i)$  are multiplied by 4, and for all even indices  $i = 2, 4, 6, \dots$  the function values  $f(x_i)$  are multiplied by 2.

For our case, for  $h = \frac{1}{4}$  we get the equation  $\frac{1}{4} = \frac{a-b}{n} = \frac{1-0}{n}$ , which gives  $n = 4$ . Thus we have  $n + 1 = 5$  nodes:  $x_0 = 0, x_1 = 1/4, x_2 = 1/2, x_3 = 3/4$  and  $x_4 = 1$ , with values  $f(x_0) = 1, f(x_1) = 4/5, f(x_2) = 2/3, f(x_3) = 4/7$  and  $f(x_4) = 1/2$ , respectively. Thus

$$I_{\text{Simp}} = \frac{h}{3} [f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + f(x_4)] = \frac{1}{12} \left( 1 + 4\frac{4}{5} + 2\frac{2}{3} + 4\frac{4}{7} + \frac{1}{2} \right) \approx 0.69325,$$

where we used Sage via the following basic command to do the final computation:

```
print(N((1/12)*(1+16/5+4/3+16/7+1/2)))
```

Let us now find the approximation by the trapezoid rule, where the corresponding nodes are the same. We have

$$I_{\text{trap}} = \frac{h}{2} [f(x_0) + 2f(x_1) + 2f(x_2) + 2f(x_3) + f(x_4)] = \frac{1}{8} \left( 1 + 2\frac{4}{5} + 2\frac{2}{3} + 2\frac{4}{7} + \frac{1}{2} \right) \approx 0.6970.$$

This result can be verified by using the program constructed in 6.B.67, for example. Hence we can type

```
def trapezoid_rule(f, a, b, n):
    step = (b - a) / n
    valsf = [f(a + i * step) for i in range(n + 1)]
    return (step / 2) * (valsf[0] + 2 * sum(valsf[1:n]) + valsf[-1])
n=4; a=0; b=1; f(x)=1/(1+x)
print(trapezoid_rule(f, a, b, n).n())
```

Running this block Sage prints out the number 0.697023809523809.

The exact value of  $I$  is given by  $I = \int_0^1 \frac{1}{1+x} dx = \ln(1+1) - \ln(1+0) = \ln(2) \approx 0.69314$ , and we may summarize the results about the errors in a table:

Method	Approximation	Actual error
Trapezoidal Rule	0.6970	$ 0.693147 - 0.6970  \approx 0.0039$
Simpson's Rule	0.693253	$ 0.693147 - 0.693253  \approx 0.000106$

Thus, the approximation given by Simpson's rule is significantly closer to the exact value of  $I$ , compared to the trapezoidal rule (Simpson's rule yields a smaller error).

Similarly is treated the case with  $h = 1/2$ , which corresponds to  $n = 2$  and  $n + 1 = 3$  nodes:  $x_0 = 0$ ,  $x_1 = 1/2$ , and  $x_2 = 1$ . We leave this for practice.

**6.B.69.** Recall that Simpson's rule requires  $n$  to be an even number, as the method involves pairing subintervals to fit a quadratic polynomial. To solve this task, we will use a method similar to the one presented in 6.B.67, and therefore, we will omit extensive explanations. Here is our routine:

```
def simpson_rule(f, a, b, n):
    if n % 2 != 0:
        raise ValueError("n must be even for Simpson's rule.")

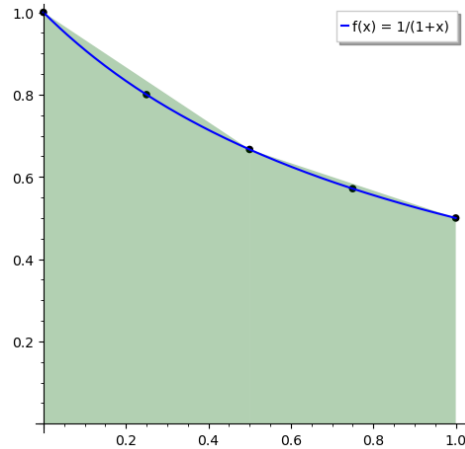
    step = (b - a) / n
    valsf = [f(a + i * step) for i in range(n + 1)]
    return (step / 3) * (valsf[0] + 4 * sum(valsf[i] for i in range(1, n, 2))
        + 2 * sum(valsf[i] for i in range(2, n-1, 2)) + valsf[-1])
```

Let us now use the function  $f(x) = 1/(1+x)$  of the previous example, with  $h = 1/4$  to check our program:

```
n = 4; a = 0; b = 1; f(x) = 1 / (1 + x)
print(simpson_rule(f, a, b, n).n())
```

This verifies our computation, as Sage's output is the number 0.69325396825396. Note that if  $n$  is odd, Sage will display the error: " $n$  must be even for Simpson's rule." You can use this program to demonstrate additional examples of Simpson's rule in your Sage editor.

For  $h = 1/4$  and the interval  $I$  given in 6.B.68, the slight overestimation of Simpson's rule demonstrated above can be illustrated as follows:



To construct this figure we used the following block in Sage:

```
f(x) = 1 / (1 + x)
a = 0; b = 1; h = 1/4
n = int((b - a) / h)
# Generate x values and corresponding y values
x_values = [a + i*h for i in range(n + 1)]
y_values = [f(x) for x in x_values]
# Initialize the plot with the original function
p = plot(f, a, b, color="blue", legend_label="f(x) = 1/(1+x)", thickness=1.5)
# Create the piecewise quadratic approximation for Simpson's rule
for i in range(0, n, 2):
    x0, x1, x2 = x_values[i], x_values[i+1], x_values[i+2]
    y0, y1, y2 = y_values[i], y_values[i+1], y_values[i+2]
    # Shade the area under the curve for this segment using a polygon
    p += polygon([(x0, 0), (x0, y0), (x2, y2), (x2, 0)], color="darkgreen", alpha=0.3, fill=True)

# Highlight the points used in Simpson's rule
p += points([(x, f(x)) for x in x_values], color="black", size=30, marker="o")
# Show the plot with a larger range for visibility
p.show(legend_loc="upper right")
```

**6.B.70.** It is easy to see that the fourth derivative of  $f(x) = \frac{1}{1+x}$  is given by  $f^{(4)}(x) = \frac{24}{(1+x)^5}$ , for all  $x \in [0, 1]$ . Recall that we can verify this computation in Sage by the cell

```
f(x)=1/(1+x); dif4=diff(f, x, 4); show(dif4)
```

Since the fourth derivative  $f^{(4)}(x)$  is decreasing for all  $x \in [0, 1]$ , the maximum value of  $|f^{(4)}(x)|$  occurs at the left endpoint of the interval, i.e., at  $x = 0$ , and equals 24. Here is a short program in Sage that can verify this claim (you may also plot the graph of  $f^{(4)}(x)$ )

```
var("x"); f_4 = 24 / (1 + x)^5
a = 0; b = 1
# Evaluate f^{(4)} at the endpoints
f_4_a = f_4.subs(x=a)
f_4_b = f_4.subs(x=b)
# Find the maximum value of |f^{(4)}(x)| on the interval
max_value = max(f_4_a, f_4_b)
# Output the results
f^{(4)}(0) = {f_4_a}, f^{(4)}(1) = {f_4_b}, Maximum |f^{(4)}(x)| on [0, 1] = {max_value}"
```

Thus we compute

$$\frac{(b-a)^5}{180n^4} \max_{x \in [0,1]} |f^{(4)}(x)| \stackrel{n=4}{=} \frac{24}{180 \cdot 256} = \frac{1}{1920} \approx 0.00052.$$

The actual error 0.000106 computed above is indeed less than the theoretical error bound 0.00052. In this way one confirms that the theoretical error bound is valid, and provides an upper limit on the actual error.

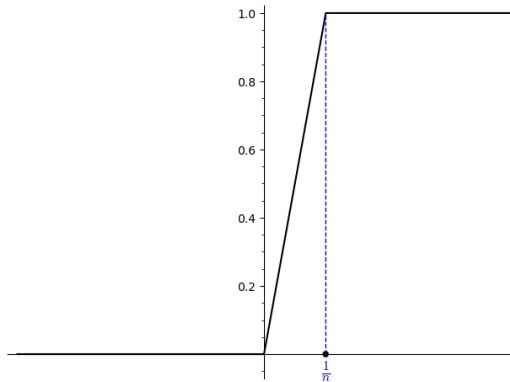
**6.C.3.** (a) A typical example is the sequence  $(f_n)$  with  $f_n(x) = x^n$ ,  $x \in [0, 1]$ . In 5.B.3 we learned that for  $0 \leq x < 1$  the sequence  $(f_n(x)) = (x^n)$  converges to 0, while  $(f_n(1))$  converges to 1. Thus, the limit function of  $(f_n)$  on  $[0, 1]$  is given by

$$f(x) = \begin{cases} 0, & \text{if } x \in [0, 1), \\ 1, & \text{if } x = 1. \end{cases}$$

Although each member  $f_n$  is continuous, it is obvious that  $f$  is not continuous at  $x_0 = 1$  (it has a jump discontinuity there). Another example is given by the sequence  $(f_n)_{n \in \mathbb{Z}_+}$  defined by

$$f_n(x) = \begin{cases} 0, & \text{if } x < 0, \\ nx, & \text{if } 0 \leq x \leq \frac{1}{n}, \\ 1, & \text{if } x > \frac{1}{n}. \end{cases}$$

The graph of  $f_n$  has the following form:



This was made in Sage by the cell

```
var("n"); n=2
def f(x):
    if -2 <x<0:
        return 0
    elif 0<=x<=1/n:
        return n*x
    elif 1/n<x<2:
        return 1
p=plot(f, (-2, 2), thickness=1.5, color="black")
p+=point([1/2, 0], size=30, color="black")
p+=line([(1/2, 0), (1/2, 1)], linestyle="--", color="blue")
p+=text(r"\frac{1}{n}", (1/2+0.01, -0.05), color="darkblue", fontsize="16")
p.show(ticks=[[], None])
```

Now, it is easy to see that  $f_n(x)$  is continuous on  $\mathbb{R}$  (notice that  $\lim_{x \rightarrow (\frac{1}{n})^-} f(x) = \lim_{x \rightarrow (\frac{1}{n})^-} nx = 1 = \lim_{x \rightarrow (\frac{1}{n})^+} f(x)$ ). However, the pointwise limit function of this sequence is the discontinuous function

$$f(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}$$

(b) Consider for instance the sequence  $(f_n(x) = e^{-nx^2})$  with  $n \in \mathbb{N}$  and  $x \in \mathbb{R}$ . For  $x = 0$  we obviously have  $f_n(0) = 1$  and  $\lim_{n \rightarrow \infty} f_n(0) = 1$ , for all  $n$ . For  $x \neq 0$  we have  $\lim_{n \rightarrow \infty} f_n(x) = 0$ , for all  $n$ . This is because of the inequality  $\frac{1}{e^t} < \frac{1}{1+t}$ , which holds for any  $t > 0$ .<sup>8</sup> This gives that

$$0 < f_n(x) = \frac{1}{e^{nx^2}} < \frac{1}{1+nx^2},$$

and our claim follows. Thus, the sequence  $(f_n)$  converges on  $\mathbb{R}$  to the function

$$f(x) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{if } x \neq 0. \end{cases}$$

<sup>8</sup>To see this, use the familiar inequality  $e^t > 1 + t$ ,  $t > 0$ .

Now it is obvious that though all members of  $(f_n)$  are differentiable functions over  $\mathbb{R}$ , the limit function  $f(x)$  is not differentiable at  $x = 0$ . Another such example is given in 6.C.6.

(c) Let us consider the sequence of functions  $(f_n : [0, 1] \rightarrow \mathbb{R})$  defined by

$$f_n(x) = \begin{cases} 0, & \text{for } x = 0, \\ n, & \text{for } 0 < x \leq \frac{1}{n}, \\ 1/x, & \text{if } \frac{1}{n} < x \leq 1, \end{cases}$$

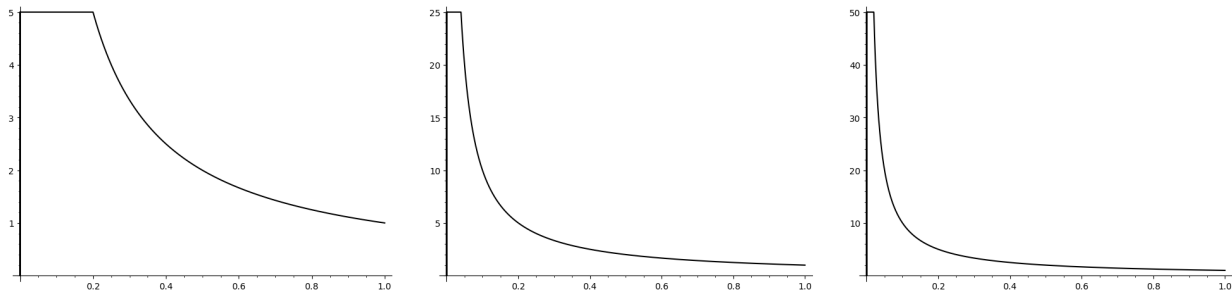
It is not hard to prove that the pointwise limit function  $f : [0, 1] \rightarrow \mathbb{R}$  of  $(f_n)_{n \in \mathbb{Z}_+}$  is given by

$$f(x) = \begin{cases} 0, & \text{if } x = 0, \\ 1/x, & \text{if } 0 < x \leq 1, \end{cases}$$

To illustrate this result graphically, we just need to use Sage and plot some members of the sequence  $(f_n)$ . This can be accomplished in several different ways. For instance, you may use a bit of programming as demonstrated in the following block:

```
var("n")
n=5
def f(x):
    if x==0:
        return 0
    elif 0<x<=1/n:
        return n
    elif 1/n<x<=1:
        return 1/x
p=plot(f, (x, 0, 1), thickness=1.5, color="black")
show(p)
```

This will produce the plot of  $f_5(x)$  and for convenience below we present the graphs of  $f_5(x)$ ,  $f_{25}(x)$  and  $f_{50}(x)$ .



Let us now prove that any member  $f_n(x)$  is Riemann integrable on  $[0, 1]$ . It is easy to see that each  $f_n(x)$  is bounded on  $[0, 1]$  (for instance, on the interval  $(\frac{1}{n}, 1]$  we see that  $\frac{1}{x}$  is bounded since  $x > \frac{1}{n}$  implies  $\frac{1}{x} < n$ ). Next we need to understand the discontinuities of the functions  $f_n, n \in \mathbb{Z}_+$ . Obviously, at  $x = 0$  we have a (jump) discontinuity,  $\lim_{x \rightarrow 0^+} f_n(x) = n \neq 0 = f_n(0)$ , for all  $n \in \mathbb{Z}_+$ . At all the other points the functions  $f_n$  are continuous: e.g., at  $x = 1/n$  we get

$$\lim_{x \rightarrow (1/n)^-} f_n(x) = \lim_{x \rightarrow (1/n)^+} n = n = f_n(1/n), \quad \lim_{x \rightarrow (1/n)^+} f_n(x) = \lim_{x \rightarrow (1/n)^+} (1/x) = \frac{1}{1/n} = n.$$

To summarize, each member  $f_n$  is bounded and has a unique discontinuity at  $x = 0$  (so a finite number of discontinuities, and thus of measure zero). Therefore, each  $f_n$  is Riemann integrable. In fact, we can easily see that

$$\int_0^1 f_n(x) dx = \int_0^{1/n} n dx + \int_{1/n}^1 \frac{1}{x} dx = 1 + \ln(n).$$

In contrast, the limit function  $f$  is not bounded on  $[0, 1]$ , and thus cannot be integrable on  $[0, 1]$ .

**6.C.5.** The sequence of functions  $(f_n)$  defined by the given graph has the form

$$f_n(x) = \begin{cases} nx, & \text{for } x \in [0, \frac{1}{n}), \\ 2 - nx, & \text{for } x \in [1/n, 2/n], \\ 0, & \text{otherwise.} \end{cases}$$

If  $x = 0$  we have  $f_n(x) = 0$  for all  $n$ , and  $\lim_{n \rightarrow \infty} f_n(0) = 0$ . Suppose now that  $x > 0$  and in particular that  $x \in (0, \frac{1}{n})$  for some  $n$ . Then, we may find a sufficiently large integer  $n_0$  such that  $x \notin [0, \frac{1}{n_0})$ , which implies again that  $\lim_{n \rightarrow \infty} f_n(x) = 0$ .

If  $x \in [\frac{1}{n}, \frac{2}{n}]$  then similarly there exists a sufficiently large integer  $n_0$  such that  $x \notin [\frac{1}{n_0}, \frac{2}{n_0}]$ , which again implies that  $\lim_{n \rightarrow \infty} f_n(x) = 0$ . Hence the claim, i.e.,  $\lim_{n \rightarrow \infty} f_n(x) = f(x) = 0$  for all  $x \geq 0$ .

For the uniform convergence, let  $x_0 = \frac{1}{2n}$ . Since  $f_n(x_0) = f_n(\frac{1}{2n}) = n \cdot \frac{1}{2n} = \frac{1}{2}$  is non-zero (regardless how large  $n$  becomes), we also get  $|f_n(\frac{1}{2n}) - f(\frac{1}{2n})| = |\frac{1}{2} - 0| = \frac{1}{2}$ . Therefore, for any  $\varepsilon < \frac{1}{2}$  does not exist  $N$  such that  $|f_n(x)| < \varepsilon$  for all  $n \geq N$ . Consequently,  $(f_n(x))$  cannot converge uniformly to  $f(x) = 0$ .

**6.C.6.** Obviously,

$$\lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \sqrt{x^2 + \frac{1}{n^2}} = \sqrt{x^2 + \lim_{n \rightarrow \infty} \frac{1}{n^2}} = \sqrt{x^2 + 0} = \sqrt{x^2} = |x|,$$

that is, the limit function is given by  $f(x) = |x|$  with  $x \in \mathbb{R}$ . Now, for given  $\varepsilon > 0$ , choose  $N$  large enough such that  $\frac{1}{N} < \varepsilon$ . Then, for any  $n \geq N$  and  $x \in \mathbb{R}$  we have

$$\begin{aligned} |f_n(x) - f(x)| &= \left| \sqrt{x^2 + \frac{1}{n^2}} - |x| \right| = \left| \sqrt{x^2 + \frac{1}{n^2}} - |x| \right| \cdot \frac{\sqrt{x^2 + \frac{1}{n^2}} + |x|}{\sqrt{x^2 + \frac{1}{n^2}} + |x|} \\ &= \frac{\frac{1}{n^2}}{\sqrt{x^2 + \frac{1}{n^2}} + |x|} \leq \frac{\frac{1}{n^2}}{\sqrt{\frac{1}{n^2}}} = \frac{1}{n} < \varepsilon. \end{aligned}$$

This proves the claim. Observe that each member  $f_n(x)$  is continuous and differentiable for all  $x \in \mathbb{R}$ , but the limit function  $f(x) = |x|$ , though continuous, is not differentiable at  $x = 0$ .

**6.C.7.** (a) On the interval  $I = (0, 1)$  consider the sequences  $(f_n)$ ,  $(g_n)$ , defined by

$$f_n(x) = \frac{1}{x}, \quad g_n(x) = \frac{1}{n}, \quad n \in \mathbb{Z}_+.$$

Obviously,  $f_n \rightarrow f = \frac{1}{x}$  and  $g_n \rightarrow 0$ , as  $n \rightarrow \infty$ , and it is easy to see that we have uniform convergence on  $(0, 1)$ . Consider the sequence  $(f_n \cdot g_n)_{n \in \mathbb{Z}_+}$  with  $f_n(x)g_n(x) = \frac{1}{nx}$  for all  $x \in (0, 1)$ . Certainly,  $\frac{1}{nx} \rightarrow 0$  as  $n \rightarrow \infty$ , hence its pointwise limit function is the zero one. We will show however that the convergence is not uniform on  $(0, 1)$ .

For uniform convergence, given  $\varepsilon > 0$  we need to find some positive integer  $N$  such that  $|\frac{1}{nx} - 0| < \varepsilon$  for all  $n \geq N$  and  $x \in (0, 1)$ . This is also written as  $\frac{1}{nx} < \varepsilon$  or equivalently,  $n > \frac{1}{\varepsilon x}$ , which must hold for all  $x \in (0, 1)$ . However, no matter how large  $n$  is chosen, there will always be some  $x \in (0, 1)$  close enough to 0 such that  $\frac{1}{nx} \geq \varepsilon$  (notice that as  $x \rightarrow 0^+$ , the expression  $1/x$  grows without bound, making  $\frac{1}{nx}$  difficult to keep below any positive  $\varepsilon$  uniformly). In particular, we see that

$$\sup_{x \in (0, 1)} |f_n(x)g_n(x) - 0| = \sup_{x \in (0, 1)} \left| \frac{1}{nx} \right| = \sup_{x \in (0, 1)} \left( \frac{1}{nx} \right) = +\infty.$$

**6.C.8.**

**6.D.4.** We compute  $T_0^3(g(x)) = 1 + \frac{x^2}{2}$ ,  $T_0^3(h(x)) = 1 - \frac{x^2}{2}$  and  $T_0^3(k(x)) = x - \frac{x^3}{3}$ , respectively. In Sage we can verify these results as usual, that is, by typing

`show(taylor(1/cos(x), x, 0, 3)), show(taylor(exp(-x**2/2), x, 0, 3)), show(taylor(sin(sin(x)), x, 0, 3)),`  
respectively.

**6.D.9.** (a) The  $x$ -axis  $y = 0$  as  $x \rightarrow \pm\infty$ . (b) The lines  $y = \ln 10$ , and  $y = x + \ln 3$ .

**6.D.13.** The function  $f(x) = -x^2/(x+1)$  with  $x \in A = \mathbb{R} \setminus \{-1\}$  is not odd, neither even, nor periodic. Its range is the union  $(-\infty, 0] \cup [4, +\infty)$ . Moreover, the point  $x_0 = -1$  is an improper point, i.e.,  $f$  is not defined there and so it has a single discontinuity with

$$\lim_{x \rightarrow -1^+} f(x) = -\infty, \quad \lim_{x \rightarrow -1^-} f(x) = +\infty.$$

The function intersects the  $x$  axis only at the origin. It is positive for  $x < -1$  and not positive for  $x > -1$ . It can be shown easily that

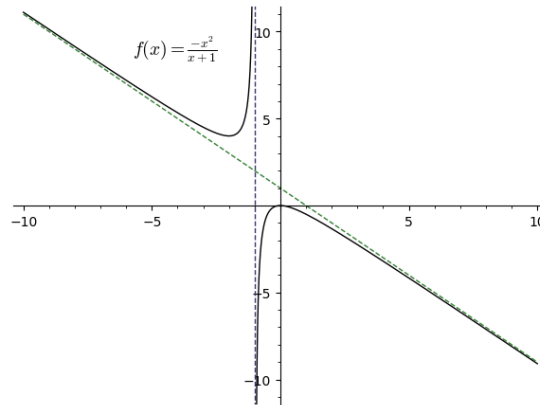
$$\lim_{x \rightarrow -\infty} f(x) = +\infty, \quad \lim_{x \rightarrow +\infty} f(x) = -\infty, \quad f'(x) = -\frac{x^2 + 2x}{(x+1)^2}, \quad f''(x) = -\frac{2}{(x+1)^3}, \quad x \in \mathbb{R} \setminus \{-1\}.$$

Therefore,  $f$  is increasing on the intervals  $[-2, -1)$ ,  $(-1, 0]$  and decreasing on the intervals  $(-\infty, -2]$ ,  $[0, +\infty)$ , see also a sketch of its graph below. The function  $f$  has two stationary points  $x_1 = 0$  and  $x_2 = -2$ . We leave the characterization of these critical points to the reader. Moreover,  $f$  is convex on the interval  $(-\infty, -1)$  and concave on the interval  $(-1, +\infty)$ .



However, it does not have a point of inflection. The line  $x = -1$  is a vertical asymptote, the inclined asymptote at  $\pm\infty$  is the line  $y = -x + 1$ . For example,  $f(-3) = 9/2$ ,  $f'(-3) = -3/4$ ,  $f(1) = -1/2$ ,  $f'(1) = -3/4$ .

To graph of  $f$  together with its asymptotes is given below:



To obtain this illustration one may use the block

```
f(x) = (-x^2) / (x+1)
p=plot(f(x), x, -10, 10, exclude=[-1], ymin=-11, ymax=11, color="black", figsize=6)
p+=text(r"$f(x) = \frac{-x^2}{x+1}$", (-4.1, 9), fontsize=14, color="black")
p+=line([(-1, -12), (-1, 12)], rgbcolor=(0.2,0.2,0.5), linestyle="--")
p+=line([(-10, 11), (10, -9)], rgbcolor=(0.2,0.5,0.2), linestyle="--"); show(p)
```

Notice in the second line we have used the option `exclude = [-1]`, in order to exclude the point where  $f$  is not defined.

**6.D.14.** The function is defined on  $A = \mathbb{R} \setminus \{1\}$  and is everywhere continuous on  $A$ . It is not odd, neither even nor periodic. The points of intersection of the graph of  $f$  with the axes are the points  $[1 - \sqrt[3]{2}, 0]$  and  $[0, -1]$ . At  $x_0 = 1$ , the function has a discontinuity of the second kind and its range is  $\mathbb{R}$ , which follows from the limits

$$\lim_{x \rightarrow 1^-} f(x) = -\infty, \quad \lim_{x \rightarrow 1^+} f(x) = +\infty, \quad \lim_{x \rightarrow \pm\infty} f(x) = +\infty.$$

After the arrangement

$$f(x) = (x-1)^2 + \frac{2}{x-1}, \quad x \in \mathbb{R} \setminus \{1\},$$

it is not difficult to compute

$$f'(x) = 2 \frac{(x-1)^3 - 1}{(x-1)^2}, \quad x \in \mathbb{R} \setminus \{1\},$$

$$f''(x) = 2 \frac{(x-1)^3 + 2}{(x-1)^3}, \quad x \in \mathbb{R} \setminus \{1\}.$$

The only stationary point is  $x_1 = 2$ . The function  $f$  is increasing on the interval  $[2, +\infty)$ , decreasing on the intervals  $(-\infty, 1)$ ,  $(1, 2]$ . Hence at the point  $x_1$  it attains the local minimum  $y_1 = 3$ . It is convex on the intervals  $(-\infty, 1 - \sqrt[3]{2})$ ,  $(1, +\infty)$  and concave on the intervals  $(1 - \sqrt[3]{2}, 1)$ . The point  $x_2 = 1 - \sqrt[3]{2}$  is a point of inflection. The line  $x = 1$  is a horizontal asymptote. The function does not have any inclined asymptotes.

**6.D.15.** The function is defined on  $\mathbb{R}$  and is everywhere continuous. It is not odd, even nor periodic. It attains positive values on the positive half-axis, negative values on the negative half-axis. The point of intersection of the graph of  $f$  with the axes is only at the point  $[0, 0]$ . The derivative is:

$$f'(x) = \frac{e^{-x}}{3\sqrt[3]{x^2}} - \sqrt[3]{x} e^{-x}, \quad x \in \mathbb{R} \setminus \{0\}, \quad f'(0) = +\infty,$$

$$f''(x) = \sqrt[3]{x} e^{-x} - \frac{2e^{-x}}{3\sqrt[3]{x^2}} - \frac{2e^{-x}}{9\sqrt[3]{x^5}}, \quad x \in \mathbb{R} \setminus \{0\}.$$

The only zero point of the first derivative is the point  $x_0 = 1/3$ . The function  $f$  is increasing on the interval  $(-\infty, 1/3]$  and decreasing on the interval  $[1/3, +\infty)$ . Hence at the point  $x_0$ , it has an absolute maximum  $y_0 = 1/\sqrt[3]{3e}$ . Since  $\lim_{x \rightarrow -\infty} f(x) = -\infty$ , its range is  $(-\infty, y_0]$ . The points of inflection are

$$x_1 = \frac{1-\sqrt{3}}{3}, \quad x_2 = 0, \quad x_3 = \frac{1+\sqrt{3}}{3}.$$

It is convex on the intervals  $(x_1, x_2)$  and  $(x_3, +\infty)$ , concave on the intervals  $(-\infty, x_1)$ ,  $(x_2, x_3)$ . The only asymptote is the line  $y = 0$  at  $+\infty$ , i.e.  $\lim_{x \rightarrow +\infty} f(x) = 0$ .

**6.D.17.** Recall by 6.A.45 that the slope of the cycloid

$$x(t) = t - \sin(t), \quad y(t) = 1 - \cos(t), \quad t \in [0, 2\pi]$$

is given by

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{\sin(t)}{1 - \cos(t)}.$$

We see that  $\frac{dy}{dx} = 0$  if and only if  $\frac{dy}{dt} = 0$ , that is,  $\sin(t) = 0$ . This gives  $t = \pi$  where we have a parallel tangent line. The vertical tangents appear at the cusps of the cycloid, that is, at the point  $t = 0$  or  $t = 2\pi$ . These values of  $t$  solve the equation  $\frac{dx}{dt} = 0$ , that is,  $\cos(t) = 1$ . To calculate the slopes at these points we can use limits:

$$k_0 = \lim_{t \rightarrow 0^+} \frac{dy}{dx} = \lim_{t \rightarrow 0^+} \frac{\sin(t)}{1 - \cos(t)} = +\infty, \quad k_{2\pi} = \lim_{t \rightarrow 2\pi^-} \frac{dy}{dx} = \lim_{t \rightarrow 2\pi^-} \frac{\sin(t)}{1 - \cos(t)} = -\infty.$$

Thus the tangents at  $t = 0, 2\pi$  are vertical.

**6.D.22.** Here one should simply give the following block:

```
show(integral(x^4+e^x+5*ln(x), x)); show(integral(sqrt(x)*(1+x^3), x)); show(integral(x/sqrt(x+1), x))
```

Check yourself Sage's output.

**6.D.23.** (a)  $F(x) = \frac{8}{15} x \sqrt[8]{x^7}$ ;

(b)  $G(x) = \frac{4^x}{\ln 4} + 2 \frac{6^x}{\ln 6} + \frac{9^x}{\ln 9}$ ;

(c)  $H(x) = \frac{\arcsin(x)}{2}$ ;

(d)  $K(x) = \ln(1 + \sin(x))$ .

**6.D.24.** A primitive function is given by  $F(x) = e^x + 3 \arcsin\left(\frac{x}{2}\right)$ .

**6.D.25.** Set  $t = \ln(x)$  with  $dt = \frac{1}{x} dx$ . Then we see that

$$\int \frac{dx}{x(\ln(x))^2 + 2025x} = \int \frac{dx}{x((\ln(x))^2 + 2025)} = \int \frac{dt}{t^2 + 45^2} = \frac{1}{45} \arctan\left(\frac{t}{45}\right) + C = \frac{1}{45} \arctan\left(\frac{\ln(x)}{45}\right) + C,$$

for some constant  $C \in \mathbb{R}$ . If you like to verify the result in Sage just type

```
show(integral(1/(x*((ln(x))^2+2025)), x))
```

**6.D.26.** Set  $t = e^x$  such that  $dt = e^x dx = t dx$ , that is,  $dx = \frac{1}{t} dt$ . Then

$$\int \frac{e^x}{e^x + 2024} dx = \int \left( \frac{t}{t + 2024} \cdot \frac{1}{t} \right) dt = \int \frac{dt}{t + 2024} = \ln|t + 2024| + C = \ln(t + 2024) + C = \ln(e^x + 2024) + C.$$

**6.D.34.** To be written.

**6.D.35.** The answer is  $\frac{3}{2} \ln(x^2 + 4x + 8) - \frac{1}{2} \operatorname{arctg} \frac{x+2}{2} + C$ , for some  $C \in \mathbb{R}$ .

**6.D.36.** The answer is  $\frac{4}{3\sqrt{3}} \operatorname{arctg} \frac{2x+1}{\sqrt{3}} + \frac{2x+1}{3(x^2+x+1)} + C$ , with  $C \in \mathbb{R}$ .

**6.D.37.** The answer is  $\frac{1}{6} \ln \frac{(x+1)^2}{x^2-x+1} + \frac{\sqrt{3}}{3} \operatorname{arctg} \frac{2x-1}{\sqrt{3}} + C$ , with  $C \in \mathbb{R}$ .

**6.D.41.** We compute  $A = \frac{1}{2} \ln \left( \frac{2 + \ln 2}{2 - \ln 2} \right)$ ,  $B = -\frac{1}{6} - \frac{2}{9} \ln(2)$ , and  $C = 2/3$ .

**6.D.48.** The answer is

$$\sum_{i=0}^{\infty} (-1)^i \frac{2^{2n-1}}{(2n)!} x^{2n},$$

which converges for all real  $x$ .

**6.D.49.** The answer is

$$\sum_{n=1}^{\infty} (-1)^{n+1} \frac{2^{2n-1}}{(2n)!} x^{2n},$$

which converges for all real  $x$ .

**6.D.50.** The answer is

$$f(x) = \sum_{n=1}^{\infty} \frac{3(-1)^{n+1}}{n} x^n,$$

which converges for all  $x \in (-1, 1]$ .

**6.D.51.** One should first realize that we are expanding the function  $f(x) = \frac{1}{2} \ln(x)$ . Thus, we get

$$f(x) = \sum_{i=0}^{\infty} (-1)^{i+1} \frac{1}{2i} (x-1)^i,$$

which converges on the interval  $(0, 2]$ .

**6.D.53.** Recall from Chapter 5 (see for example 5.D.21) that  $\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$ , which makes sense for any  $x \in \mathbb{R}$ . Thus, replacing  $x$  by  $\sqrt{x}$  we get

$$\cos(\sqrt{x}) = 1 - \frac{(\sqrt{x})^2}{2!} + \frac{(\sqrt{x})^4}{4!} - \frac{(\sqrt{x})^6}{6!} + \dots = 1 - \frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \dots.$$

We can now approximate the integral at hand, as follows:

$$I = \int_0^1 \left(1 - \frac{x}{2!} + \frac{x^2}{4!} - \frac{x^3}{6!} + \dots\right) dx = \left[x - \frac{x^2}{2 \cdot 2!} + \frac{x^3}{3 \cdot 4!} - \frac{x^4}{4 \cdot 6!} + \dots\right]_0^1 = 1 - \frac{1}{2 \cdot 2!} + \frac{1}{3 \cdot 4!} - \frac{1}{4 \cdot 6!} + \dots$$

Hence we deduce that  $I = \int_0^1 \cos(\sqrt{x}) dx \approx 1 - \frac{1}{2 \cdot 2!} + \frac{1}{3 \cdot 4!} - \frac{1}{4 \cdot 6!} = 1 - \frac{1}{4} + \frac{1}{72} - \frac{1}{2880} = 733/960 \approx 0.7635416$ .

In Sage the real value of  $I$  is obtained by the cell

```
integral(cos(sqrt(x)), x, 0, 1)
N(integral(cos(sqrt(x)), x, 0, 1))
```

Here the first commands gives the answer  $2 \cos(1) + 2 \sin(1) - 2 \approx$  and the second prints out its decimal expression, 0.763546581352073. Hence the estimation based on the Maclaurin series has very small error.

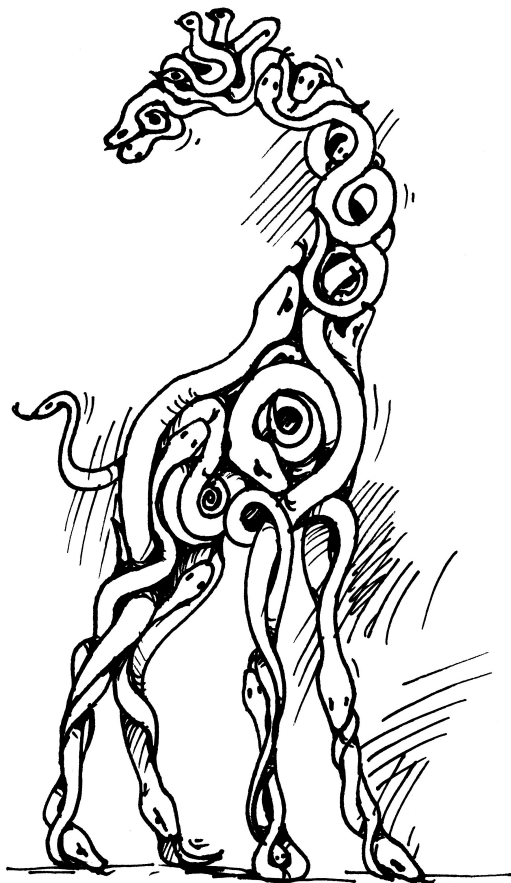
**6.D.56.** The error belongs to the interval  $(0, 1/200)$ .

**6.D.57.** We find that

$$0 < \int_1^2 \frac{\cos^{10}(x)}{10} \ln(x) dx < \frac{1}{10}, \quad \int_1^2 x \ln x dx = \ln 4 - \frac{3}{4}.$$

## Continuous tools for modelling

*How do we manage non-linear objects?  
– mainly by linear tools again...*



A. Fourier series

*If we want to understand three-dimensional objects, we often use (one or more) two-dimensional plane projections of them. The orthogonal projections are special in providing the closest images of the points of the objects in the chosen plane.*

*Similarly, we can understand complicated functions in terms of simpler ones. We consider their projections into the (real) vector space generated by those chosen functions. Recall from Chapter 2 that the orthogonal projections were easily computed in terms of inner products. Now we do the same for the infinite dimensional spaces of functions.*

In this chapter, we mainly deal with applications of the tools of differential and integral calculus. We consider a variety of problems related to functions of one real variable.

The tools and procedures are similar to the ones shown in Chapter 3, i.e. we consider linear combinations of selected generators and linear transformations.

This chapter serves also as a useful consolidation of background material before considering functions of several variables, differential equations, and the calculus of variations.

We begin by asking how to approximate a given function by linear combinations from a given set of generators. Approximation considerations lead to the general concept of distance. We illustrate the concepts on rudiments of the Fourier series. Our intuition from the Euclidean spaces of low dimensions is extended to infinite dimensional spaces, particularly the concept of orthogonal projections.

The next part of this chapter focuses on integral operators. These are linear mappings on functions which are defined in terms of integrals. Especially, we pay attention to convolutions and Fourier analysis. Throughout all these considerations, we work with real or complex valued functions of one variable.

Only then do we introduce the elements of the theory of metric spaces. This should enlighten the concepts of convergence and approximation on infinite dimensional spaces of functions. It will also cover our needs in analysis on Euclidean spaces  $\mathbb{R}^n$  in the next chapter.

### 1. Fourier series

**7.1.1. Spaces of functions.** As usual, we begin by choosing appropriate sets of functions to use. We want enough functions so that our models can conveniently be applied in practice. At the same time, the functions must be sufficiently “smooth” so that we can integrate and differentiate as needed.

All functions are defined on an interval  $I = [a, b] \subset \mathbb{R}$ , where  $a < b$ . The interval may be bounded, (i.e., both  $a$  and  $b$  are finite), or unbounded (i.e., either  $a = -\infty$ , or  $b = +\infty$ , or both).

In this case the inner product mimics the product of scalars and provides the necessary tool to calculate the corresponding projections. The simplest way to define an inner product on appropriate vector spaces of functions takes the form

$$\langle f, g \rangle = \int_a^b f(x)g(x) dx.$$

We refer to this inner product as  $L_2$  and denote the corresponding norm by  $\| \cdot \|_2$ , see 7.1.1-7.1.3 for further details and its extension to complex functions.

**7.A.1. Orthogonal systems of functions.** Consider the subspace  $W = \langle f_1, f_2 \rangle$  of the space of real-valued functions defined on the interval  $[1, 2]$ , generated by the functions  $f_1(x) = 1/x$  and  $f_2(x) = x^2$ , endowed with the  $L_2$  product.

- (a) Complete the function  $f_1$  to an orthogonal basis of  $W$ .
- (b) Determine the orthogonal projection of  $f(x) = x$  onto  $W$  and compute the distance of  $f$  from  $W$ .

**Solution.** (a) The vector space  $W$  is generated by two linearly independent functions, thus its dimension is 2. All vectors belonging to  $W$  are of the form  $a \cdot f_1(x) + b \cdot f_2(x)$  for some  $a, b \in \mathbb{R}$ . According to the Gram-Schmidt process, we seek for a vector of the form  $h = x^2 + k \cdot \frac{1}{x}$ ,  $k \in \mathbb{R}$ , subject to the orthogonality condition  $0 = \langle \frac{1}{x}, x^2 + k \cdot \frac{1}{x} \rangle$ , which gives

$$k = -\frac{\langle \frac{1}{x}, x^2 \rangle}{\langle \frac{1}{x}, \frac{1}{x} \rangle} = -\frac{\int_1^2 \frac{1}{x} \cdot x^2 dx}{\int_1^2 \frac{1}{x} \cdot \frac{1}{x} dx} = -3.$$

Hence the requested orthogonal basis consists of the functions  $\frac{1}{x}$  and  $x^2 - \frac{3}{x}$ .

(b) The projection of the function  $f(x) = x$  onto  $W$  has the form (see ....)

$$\begin{aligned} p_x &= \frac{\langle x, \frac{1}{x} \rangle}{\langle \frac{1}{x}, \frac{1}{x} \rangle} \cdot \frac{1}{x} + \frac{\langle x, x^2 - \frac{3}{x} \rangle}{\langle x^2 - \frac{3}{x}, x^2 - \frac{3}{x} \rangle} \cdot (x^2 - \frac{3}{x}) \\ &= \frac{2}{x} + \frac{15}{34} (x^2 - \frac{3}{x}). \end{aligned}$$

Let us recall that the distance of a vector from the subspace is given by the norm of the difference between this vector and its projection. In this case we get

$$\begin{aligned} \|x - p_x\|_2 &= \left( \int_1^2 \left( x - \frac{2}{x} - \frac{15}{34} \left( x^2 - \frac{3}{x} \right) \right)^2 dx \right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{408}} \approx 0.0495. \quad \square \end{aligned}$$

**7.A.2.** Let  $W$  be the the space generated by the functions  $\frac{1}{x}, \frac{1}{x^2}, \frac{1}{x^3}$ , defined on the interval  $[1, 2]$ . Assume that  $W$  is endowed with the  $L_2$  product. Complete the function  $\frac{1}{x}$  to an orthogonal basis of  $W$ . Next determine the projection of the functions  $f(x) = \frac{1}{x^4}$  and  $g(x) = x$  onto  $W$  and find their distances from  $W$ . ○

SPACES OF PIECEWISE SMOOTH FUNCTIONS

We denote by  $\mathcal{S}^0 = \mathcal{S}^0[a, b]$  the set of all piecewise continuous functions on  $I = [a, b]$  with real or complex values. Otherwise put, all functions  $f$  in  $\mathcal{S}^0 = \mathcal{S}^0[a, b]$  have only finitely many points of discontinuity on bounded intervals. Moreover,  $f$  has finite one-sided left and right limits at every point in  $[a, b]$ . In particular,  $f$  is bounded on all bounded subintervals.

For every natural number  $k \geq 1$ , we consider the set of all piecewise continuous functions  $f$  such that all their derivatives up to order  $k$  (inclusive) lie in  $\mathcal{S}^0$ . We denote this set by  $\mathcal{S}^k[a, b]$ , or briefly  $\mathcal{S}^k$ . Note that the derivatives of functions in  $\mathcal{S}^k$  need not exist at all points, but their one-sided limits must exist.

If the interval  $I$  is unbounded, we often consider only those functions with compact support. A function with compact support means that it is identically zero outside some bounded interval of the real line. For unbounded intervals, we denote by  $\mathcal{S}_c^k$  the subset of those functions in  $\mathcal{S}^k$  which have compact support.

Functions in  $\mathcal{S}^0$  are always Riemann integrable on the bounded interval  $I = [a, b]$ , with both

$$\int_a^b |f(x)| dx < \infty, \quad \text{and} \quad \int_a^b |f(x)|^2 dx < \infty.$$

Both integrals are finite for unbounded intervals if the function  $f$  has compact support.

**7.1.2. Distance between functions.** The properties of limits and derivatives ensure that  $\mathcal{S}^k$  and  $\mathcal{S}_c^k$  are vector spaces. In finite-dimensional spaces, the distance between vectors can be expressed by means of the differences of the coordinate components. In spaces of functions, we proceed analogously and utilize the absolute value of real or complex numbers and the Euclidean distance in the following way:



THE  $L_1$  DISTANCE OF FUNCTIONS

The  $L_1$ -distance between functions  $f$  and  $g$  in  $\mathcal{S}_c^0$  is defined by

$$\|f - g\|_1 = \int_a^b |f(x) - g(x)| dx.$$

If  $g = 0$ , then the distance from  $f$  to the zero function, namely  $\|f\|_1$ , is called the  $L_1$ -norm (i.e., length or size) of  $f$ .

The  $L_1$ -distance between functions  $f$  and  $g$  (when both are real valued) expresses the area enclosed by the graphs of these functions, regardless of which function takes greater values. We observe that  $\|f - g\|_1 \geq 0$ .

Since  $f$  and  $g$  are both piecewise continuous functions,  $\|f - g\|_1 = 0$  only if  $f$  and  $g$  differ in their values at most at the points of discontinuity, and hence at only finitely many points on any bounded interval. Recall that we can change

The next task involves a specific type of polynomials, known as “Chebyshev polynomials”. These are defined by  $T_n(x) = \cos(n \arccos(x))$ , or equivalently, given by

$$T_n(\cos(x)) = \cos(nx), \quad n \in \mathbb{N}.$$

Chebyshev polynomials are described by a plethora of equivalent ways, and have a distinguished role in approximation theory. Moreover, due to their simple definition are easily adapted to symbolic computations, and here we will explain how one can manipulate them via Sage.

**7.A.3. Chebyshev polynomials.** Show that  $T_n(x)$  is a polynomial for all  $n \in \mathbb{N}$ .

**Solution.** By the definition given above it is direct that

$$T_0(x) = 1, \quad T_1(x) = x.$$

Based now on the trigonometric identity

$$\cos(rz) \cos(sz) = \frac{1}{2}(\cos((r-s)z) + \cos((r+s)z)), \quad (\star)$$

we see that

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \\ &= \cos((n+1) \arccos(x)) + \cos((n-1) \arccos(x)) \\ &= 2 \cos(n \arccos(x)) \cos(\arccos(x)) = 2x T_n(x), \end{aligned}$$

that is,

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x),$$

for all positive integers  $n$ . This is the recurrent definition of Chebyshev polynomials, and the result follows.  $\square$

**7.A.4.** (a) Verify that for each interval  $I = [a, b] \subset \mathbb{R}$  and positive continuous function  $\omega$  on  $I$ , the formula

$$\langle f, g \rangle_\omega = \int_a^b f(x)g(x)\omega(x) dx$$

defines an inner product on the continuous functions on  $I$ .

(b) Choosing  $I = (-1, 1)$  and  $\omega(x) = \omega_0(x) = (1-x^2)^{-1/2}$ , deduce that the Chebyshev polynomials  $T_k(x)$ , ( $k \in \mathbb{N}$ ), form an orthogonal system of polynomials with respect to  $\langle \cdot, \cdot \rangle_{\omega_0}$ .

**Solution.** We compare the defining formula with the  $L_2$  inner product above: Consider the substitution  $x = \varphi(z)$ , where  $\varphi$  is the inverse function to  $z = \int_a^x \omega(t) dt$ . The inverse exists, since  $\omega$  is positive and so  $z$  is a strictly increasing function of  $x$ . Thus,  $dz = \omega(x)dx$  and

$$\begin{aligned} \langle f, g \rangle_\omega &= \int_a^b f(x)g(x)\omega(x) dx \\ &= \int_0^{\varphi^{-1}(b)} f(\varphi(z))g(\varphi(z)) dz. \end{aligned}$$

In particular, the “coordinate change”  $x = \varphi(z)$  identifies the vector space of continuous functions on  $I$  with the space of continuous function on another interval equipped with the  $L_2$  inner product and so  $\langle \cdot, \cdot \rangle_\omega$  is an inner product. We leave as an easy exercise to check that  $\langle \cdot, \cdot \rangle_\omega$  satisfies the properties of an inner product.

the value of any function at a finite number of points, without changing the value of the integral.

If in particular,  $f$  and  $g$  are both continuous on  $[a, b]$ , then  $\|f - g\|_1 = 0$  implies  $f(x) = g(x)$  for all  $x \in [a, b]$ . Indeed, if  $f(x_0) \neq g(x_0)$  at a point  $x_0$ ,  $a \leq x_0 \leq b$ , and if  $f$  and  $g$  are both continuous at  $x_0$ , then  $f$  and  $g$  also differ on some small neighbourhood of  $x_0$ , and this neighbourhood, in turn, contributes a non-zero value into the integral, so that then  $\|f - g\|_1 > 0$ .

If we have three functions  $f$ ,  $g$ , and  $h$ , then, of course,

$$\begin{aligned} \int_a^b |f(x) - g(x)| dx &= \int_a^b |f(x) - h(x) + h(x) - g(x)| dx \\ &\leq \int_a^b |f(x) - h(x)| dx + \int_a^b |h(x) - g(x)| dx, \end{aligned}$$

so the usual triangle inequality

$$\|f - g\|_1 \leq \|f - h\|_1 + \|h - g\|_1$$

holds. To derive this inequality, we used only the triangle inequality for the scalars; thus it is valid for functions  $f, g \in \mathcal{S}_c^0$  with complex values as well.

$\|f - g\|_1$  is not the only way to measure distance between two functions  $f$  and  $g$ . For another way:

#### THE $L_2$ -DISTANCE

The  $L_2$ -distance between functions  $f$  and  $g$  in  $\mathcal{S}_c^0$  is defined by

$$\|f - g\|_2 = \left( \int_a^b |f(x) - g(x)|^2 dx \right)^{1/2}.$$

If  $g = 0$ , then  $\|f\|_2$ , the distance from  $f$  to the zero function, is called the  $L_2$  norm of  $f$ .

Clearly  $\|f\|_2 \geq 0$ . Moreover,  $\|f\|_2 = 0$ , implies that  $f(x) = 0$  for all  $x$  except for a finite set of points in any bounded interval. As above for the  $L_1$  norm,  $\|f - g\|_2 = 0$  only if  $f$  and  $g$  differ in their values at most at the points of discontinuity, and hence at only finitely many points on any bounded interval. In particular, if  $f$  and  $g$  are both continuous for all  $x$ , then  $\|f - g\|_2 = 0$  implies  $f(x) = g(x)$  for all  $x$ .

The square of  $\|f\|_2$  for a function  $f$  is

$$\|f\|_2^2 = \int_a^b |f(x)|^2 dx$$

and it is related to the well-defined symmetric bilinear mapping of real or complex functions to scalars

$$\langle f, g \rangle = \int_a^b f(x)\overline{g(x)} dx$$

since

$$\langle f, f \rangle = \int_a^b f(x)\overline{f(x)} dx = \int_a^b |f(x)|^2 dx = \|f\|_2^2.$$

We can use therefore all the properties of inner products in unitary spaces as described in Chapter 3. In particular, the

(b) In this special case,  $\omega(x) = \frac{d}{dx}(\arccos(x))$ , and thus the above substitution yields

$$\langle T_r, T_s \rangle_\omega = \int_0^\pi \cos(rz) \cos(sz) \, dz.$$

We are dealing with improper Riemann integrals (integrating the unbounded function  $\omega$ ), but this does not cause any problem. In fact it is easy to evaluate the integral via the trigonometric formula ( $\star$ ) mentioned in 7.A.3. We see that it vanishes for all  $r \neq s$ .  $\square$

*Both Sage and Maple provide built-in functions for generating the Chebyshev polynomials. For instance, in Sage they can be accessed via the function `chebyshev_T(n, x)`, and in our notation this correspond to  $T_n(x)$ , see also below.*

**7.A.5.** (a) Use Sage and its function `chebyshev_T` to derive the first five Chebyshev polynomials. Next, plot them.

(b) Combine the recursive definition of Chebyshev polynomials with the `def` environment in Sage to present an alternative approach for integrating them into Sage.

**Solution.** (a) To write down the first five Chebyshev polynomials one can type

```
R = PolynomialRing(QQ, "x")
show([chebyshev_T(n, x) for n in [0..4]])
```

which returns the list

$$[1, x, 2x^2 - 1, 4x^3 - 3x, 8x^4 - 8x^2 + 1]$$

To plot them add the command

```
plot([chebyshev_T(n, x) for n in [0..4]])
```

(b) For this task, let us agree to call `T(n, x)` the desired routine. We can introduce it as follows:

```
def T(n, x):
    if n==0:
        return 1
    elif n==1:
        return x
    else:
        return expand(2*x*T(n-1, x) - T(n-2, x))
```

To test the routine you may type `T(3, x)` to get  $4x^3 - 3x$ , `T(4, x)` to get  $8x^4 - 8x^2 + 1$ , etc. Or you can directly compare the formula of `T(n, x)` with that of `chebyshev_T(n, x)`, for some  $n$ , by typing

```
bool(T(n, x)==chebyshev_T(n, x) for n in [0..10])
```

In this case Sage's answer is `True`.  $\square$

**7.A.6.** A good programmer is often interested in the efficiency of algorithms and programs. For example, you may like to check the efficiency of the routine in Sage presented in 7.A.5, which can be done via the command `timeit`. Hence, one may compare the execution time of the commands `chebyshev_T(n, x)` and `T(n, x)`, for certain values of  $n$ . To test the case  $n = 3$ , for example, add the cell



inner product satisfies both linearity in the first argument and the Hermitian symmetry

$$\langle f, g \rangle = \overline{\langle g, f \rangle}.$$

It is a symmetric bilinear mapping in the real case.

In the sequel, the  $L_2$  distance will be most important. Thus, to simplify notation, the norm symbol  $\| \cdot \|$  without subscript will mean  $\| \cdot \|_2$  in the subsequent paragraphs.

**7.1.3. Orthogonality.** In Chapters 2 and 3, we dealt with finite-dimensional real or complex vector spaces. Most properties derived there concerned pairs or finite sets of vectors. Now, we can do just the same with functions. We restrict our definition of the inner product to any vector subspace generated by only finitely many functions  $f_1, \dots, f_k$  (over real or complex numbers, according to our need). We again obtain a well-defined inner product on this finite-dimensional vector subspace and so all our considerations from the finite dimensional linear algebra apply again.

As an example, consider the monomial functions  $f_i(x) = x^i, i = 0, \dots, k$ . In  $\mathcal{S}^0$ , these generate the  $(k + 1)$ -dimensional vector subspace  $\mathbb{R}_k[x]$  of all polynomials of degree at most  $k$ . The inner product of two such polynomials is given by integration. Every polynomial of degree at most  $k$  is uniquely expressed as a linear combination of the generators  $f_0, \dots, f_k$ . Moreover, if we can arrange the choice of generators to satisfy

$$(1) \quad \langle f_i, f_j \rangle = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j, \end{cases}$$

then the computations become much easier than they would otherwise.

Recall the Gram–Schmidt orthogonalization procedure, see 2.3.21. This procedure transforms any system of linearly independent generators  $f_i$  into new (again linearly independent) orthogonal generators  $g_i$  of the same subspace, i.e.  $\langle g_i, g_j \rangle = 0$  for all  $i \neq j$ . We can calculate them step by step. Put  $g_1 = f_1$ , and

$$g_{\ell+1} = f_{\ell+1} + a_1 g_1 + \dots + a_\ell g_\ell, \quad a_i = -\frac{\langle f_{\ell+1}, g_i \rangle}{\|g_i\|^2}$$

for  $\ell \geq 1$ .

To illustrate, we apply this procedure to the three polynomials  $1, x, x^2$  on the interval  $[-1, 1]$ . Put  $g_1 = 1$ , and generate the sequence

$$\begin{aligned} g_1 &= 1 \\ g_2 &= x - \frac{1}{\|g_1\|^2} \left( \int_{-1}^1 x \cdot 1 \, dx \right) \cdot g_1 = x - 0 = x \\ g_3 &= x^2 - \frac{1}{\|g_1\|^2} \left( \int_{-1}^1 x^2 \cdot 1 \, dx \right) \cdot g_1 - \\ &\quad \frac{1}{\|g_2\|^2} \left( \int_{-1}^1 x^2 \cdot x \, dx \right) \cdot g_2 = x^2 - \frac{1}{3}. \end{aligned}$$

```
print(timeit('chebyshev_T(3,x)', number=20))
print(timeit('T(3,x)', number=20))
```

Here the option `number=20` restricts the number of loops. The output appears as follows:

```
20 loops, best of 3: 81.2 μs per loop
20 loops, best of 3: 18 μs per loop
```

and what one looks for is the execution time, which here is measured in microseconds ( $\mu s$ ).<sup>1</sup> Note that each time the above two commands are executed, Sage returns slightly different execution times.

**7.A.7.** Show that the choice of the weight function  $\omega(x) = e^{-x}$  and the interval  $I = [0, \infty)$  in Problem 7.A.3, leads to an inner product for which the *Laguerre polynomials*

$$L_n(x) = \sum_{k=0}^n \binom{n}{k} \frac{(-1)^k}{k!} x^k$$

form an orthonormal system. ○

**7.A.8.** Check that the orthonormal systems obtained in the previous two examples coincide with the result of the corresponding Gram-Schmidt orthogonalisation procedure applied to the system  $1, x, x^2, \dots, x^n, \dots$ , using the inner products  $\langle \cdot, \cdot \rangle_\omega$ , possibly only up to signs. ○

**7.A.9.** Let  $\mathcal{S}^0([-\pi, \pi])$  be the (infinite-dimensional) vector space of  $\mathbb{F}$ -valued piecewise continuous functions defined on  $[-\pi, \pi]$ , where  $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ , and let us denote by  $\bar{g}$  the complex conjugate of a function  $g \in \mathcal{S}^0([-\pi, \pi])$ . Prove that for any  $f, g \in \mathcal{S}^0([-\pi, \pi])$  the rule

$$(f, g) := \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx$$

defines an inner product. Next show that the system of functions

$$\left\{ \frac{\sqrt{2}}{2}, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots : n \in \mathbb{Z}_+ \right\}$$

is  $(\cdot, \cdot)$ -orthonormal. ○

The orthonormal system of functions discussed in 7.A.9 is an orthonormal version of the so-called “Fourier orthogonal system”, introduced in 7.1.6. In fact, a key point in 7.A.9 is the “periodicity” of the functions forming the orthonormal basis (each having period  $2\pi$ ).

Periodic functions occur frequently in engineering problems, and usually have a complicated form. Thus, it is often desirable to have a presentation of such periodic functions in terms of the simpler functions of sine and cosine. Building an orthogonal system of the periodic functions  $\sin(nx)$  and  $\cos(nx)$  leads to the classical concept of the Fourier series

<sup>1</sup>For the command `timeit` Sage’s output can include also the units nanoseconds (ns) and millisecond (ms). Recall that nanoseconds provide the smallest unit of the three, followed by microseconds and then milliseconds.

The corresponding orthogonal basis of the space  $\mathbb{R}_2[x]$  of all polynomials of degree less than three on the interval  $[-1, 1]$  is  $1, x, x^2 - 1/3$ . Rescaling by appropriate numbers so that the basis elements all have length 1, yields the orthonormal basis

$$h_1 = \sqrt{\frac{1}{2}}, \quad h_2 = \sqrt{\frac{3}{2}}x, \quad h_3 = \frac{1}{2}\sqrt{\frac{5}{2}}(3x^2 - 1).$$

For example,  $h_1 = g_1 / \|g_1\|$  and

$$\|g_1\|^2 = \int_{-1}^1 1^2 dx = 2.$$

We could easily continue this procedure in order to find orthonormal generators of  $\mathbb{R}_k[x]$ . The resulting polynomials are called *Legendre polynomials*.

Considering all Legendre polynomials  $h_i, i = 0, \dots$ , we have an infinite orthonormal set of generators such that polynomials of all degrees are uniquely expressed as their finite linear combinations.

**7.1.4. Orthogonal systems of functions.** Generalizing the latter example, suppose we have three polynomials  $h_1, h_2, h_3$  forming an orthonormal set. For any polynomial  $h$ , we can put



$$H = \langle h, h_1 \rangle h_1 + \langle h, h_2 \rangle h_2 + \langle h, h_3 \rangle h_3.$$

We claim that  $H$  is the (unique) polynomial which minimizes the  $L_2$ -distance  $\|h - H\|$ . See 3.4.3.

The coefficients for the best approximation of a given function by a function from a selected subspace are obtained by the integration introduced in the definition of the inner product.

This example of computing the best approximation of  $H$  by a linear combination of the given orthonormal generators suggests the following generalization:

**ORTHOGONAL SYSTEMS OF FUNCTIONS**

Every (at most) countable system of linearly independent functions in  $\mathcal{S}_c^0[a, b]$  such that the inner product of each pair of distinct functions is zero is called an *orthogonal system of functions*. Moreover, if the norm  $\|f\|_2 = 1$  for all  $f$  in an orthogonal system, we talk about an *orthonormal system of functions*.

Notice, all continuous functions have got compact supports on a finite interval  $[a, b]$ .

Consider an orthogonal system of functions  $f_n \in \mathcal{S}^0[a, b]$  and suppose that for (real or complex) constants  $c_n$ , the series

$$F(x) = \sum_{n=0}^{\infty} c_n f_n(x)$$

converges uniformly on a finite interval  $[a, b]$ . Notice that the limit function  $F(x)$  does not need to belong to  $\mathcal{S}^0[a, b]$ , but this is not our concern now.



for a  $2L$ -periodic function  $f$  that is integrable over an interval of length  $2L$ , say  $[-L, L]$ .<sup>2</sup> These are series of the form

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right) \right),$$

where  $a_n$  and  $b_n$  are the Fourier coefficients, given by

$$a_n = \frac{1}{L} \int_{-L}^L f(x) \cos\left(\frac{n\pi x}{L}\right) dx, \quad n = 0, 1, \dots,$$

$$b_n = \frac{1}{L} \int_{-L}^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx, \quad n = 1, 2, \dots,$$

respectively.

In many cases the period is  $2\pi$ , i.e.,  $L = \pi$ , but we are also interested in functions with arbitrary periods, which, in practice (applications), is the most common scenario. For instance, if  $f$  has period  $T$  then it may be defined on any interval of the form  $[0, T]$  or  $[-T/2, T/2]$ , i.e., of length  $T$ . Defining  $f$  on intervals of the form  $[-\pi, \pi]$  or  $[0, 2\pi]$  is useful however, since it simplifies the calculations and aligns with the standard form of the Fourier series, see 7.1.6. For the purpose of Fourier series the function  $f$  may be defined even on the open interval  $(-\pi, \pi)$ , where the key requirement is that  $f$  must be periodic of period  $2\pi$ , see for example the so-called ‘‘Heaviside function’’ in 7.1.9 (also known as the ‘‘square wave function’’).

Below, our goal is to illustrate all of this with a series of examples, most of which are implemented using Sage as well. For convenience, next we will denote by  $F_k(x)$  ( $k > 0$ ) the  $k$ th-partial sum of a given Fourier series  $F(x)$ .

**7.A.10.** Consider the  $2\pi$ -periodic extension of the function  $f(x) = x$ , with  $x \in (-\pi, \pi)$ . Describe the corresponding Fourier series  $F(x)$  and next confirm your answer via Sage. Moreover, confirm that as the number of terms in the series increases, the approximation improves significantly, except in a small neighbourhood around the discontinuity point. (Hint: For example, use Sage to plot the partial sums  $F_5(x)$ ,  $F_{10}(x)$ ,  $F_{20}(x)$  and  $F_{100}(x)$ ).

**Solution.** We see that  $f(-x) = -x = -f(x)$  for all  $x \in (-\pi, \pi)$ , which means that  $f$  is an odd function. Thus, the coefficients  $a_n$  vanish for all  $n = 0, 1, \dots$ , and the Fourier series of  $f$  is a sine series, that is,  $F(x) = \sum_{n=1}^{\infty} b_n \sin(nx)$ . On the other hand, the function  $g(x) = x \sin(nx)$  is even, i.e.,  $g(-x) = g(x)$  for all  $x \in (-\pi, \pi)$ . Hence, according to the statement in 6.B.35 we have

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} x \sin(nx) dx = \frac{2}{\pi} \int_0^{\pi} x \sin(nx) dx.$$

<sup>2</sup>The Fourier series are named in honour of the French mathematician and physicist Jean B. J. Fourier, in recognition of his seminal 1807 work ‘‘On the Propagation of Heat in Solid Bodies’’, which focused on the issue of heat conduction. Fourier introduced the concept of analyzing periodic functions using trigonometric series, a method that remains fundamental to mathematical physics and has numerous critical applications in engineering.

By uniform convergence, the inner product  $\langle F, f_n \rangle$  can be expressed in terms of the particular summands (see the corollary 6.3.7), obtaining

$$\langle F, f_n \rangle = \sum_{m=0}^{\infty} c_m \int_a^b f_m(x) \overline{f_n(x)} dx = c_n \|f_n\|^2,$$

since each term in the sum is 0 except when  $m = n$ . Exactly as in the example above, each finite sum  $\sum_{n=0}^k c_n f_n(x)$  is the best approximation of the function  $F(x)$  among the linear combinations of the first  $k + 1$  functions  $f_n$  in the orthogonal system.

Actually, we can generalize the definition further to any vector space of functions with an inner product. See the exercise 7.A.4 for such an example. For the sake of simplicity we confine ourselves to the  $L_2$  distance, but the reader can check that the proofs work in general.

We extend our results from finite-dimensional spaces to infinite dimensional ones. Instead of finite linear combinations of base vectors, we have infinite series of pairwise orthogonal functions. The following theorem gives us a transparent and very general answer to the question as to how well the partial sums of such a series can approximate a given function:

**7.1.5. Theorem.** Let  $f_n$ ,  $n = 1, 2, \dots$ , be an orthogonal sequence of (real or complex) functions in  $\mathcal{S}^0[a, b]$  and let  $g \in \mathcal{S}^0[a, b]$  be an arbitrary function. Put

$$c_n = \|f_n\|^{-2} \int_a^b g(x) \overline{f_n(x)} dx.$$

Then

(1) For any fixed  $n \in \mathbb{N}$ , the expression which has the least  $L_2$ -distance from  $g$  among all linear combinations of functions  $f_1, \dots, f_n$  is

$$h_n = \sum_{i=1}^n c_i f_i(x).$$

(2) The series  $\sum_{n=1}^{\infty} |c_n|^2 \|f_n\|^2$  always converges, and moreover

$$\sum_{n=1}^{\infty} |c_n|^2 \|f_n\|^2 \leq \|g\|^2.$$

(3) The equality  $\sum_{n=1}^{\infty} |c_n|^2 \|f_n\|^2 = \|g\|^2$  holds if and only if  $\lim_{n \rightarrow \infty} \|g - h_n\| = 0$ .

Before presenting the proof, we consider the meaning of the individual statements of this theorem. Since we are working with an arbitrarily chosen orthogonal system of functions, we cannot expect that all functions can be approximated by linear combinations of the functions  $f_i$ .

For instance, if we consider the case of Legendre orthogonal polynomials on the interval  $[-1, 1]$  and restrict ourselves to even degrees only, surely we can approximate only even

Keeping in mind the relations  $\cos(n\pi) = (-1)^n$  and  $\sin(n\pi) = 0$ , integration by parts gives

$$\begin{aligned} b_n &= \frac{2}{\pi} \int_0^\pi x \sin(nx) dx = \frac{2}{\pi} \left[ -\frac{x \cos(nx)}{n} + \frac{\sin(nx)}{n^2} \right]_0^\pi \\ &= \frac{2}{\pi} \cdot \frac{-\pi}{n} (-1)^n = \frac{2}{n} (-1)^{n+1}. \end{aligned}$$

Therefore, the Fourier series of  $f$  has the form

$$\begin{aligned} F(x) &= \sum_{n=1}^\infty \frac{2(-1)^{n+1}}{n} \sin(nx) = 2 \sin(x) - \sin(2x) \\ &\quad + \frac{2}{3} \sin(3x) - \frac{1}{2} \sin(4x) + \frac{2}{5} \sin(5x) - \dots \end{aligned}$$

In Sage it suffices to compute the Fourier coefficients  $a_n$  and  $b_n$ , which can be easily done by the block

```
f=lambda x:x; f=piecewise([( -pi, pi), f])
L=pi; n=var("n")
an=(1/L)*integral(x*cos(n*pi*x/L), x, -pi, pi)
show("a_n=", an)
bn=(1/L)*integral(x*sin(n*pi*x/L), x, -pi, pi)
show("b_n=", bn.expand())
```

Sage prints out the following expressions

$$a_n=0, \quad b_n = -\frac{2 \cos(\pi n)}{n} + \frac{2 \sin(\pi n)}{\pi n^2},$$

which, with a bit of effort, confirm the presented values of  $a_n, b_n$ , and consequently the given expression of  $F(x)$ .

In Sage the description of the  $k$ th-partial sum of the Fourier series corresponding to a periodic function  $f$ , relies on the syntax `f.fourier_series_partial_sum(k, L)`, where  $L$  denotes the half-period of  $f$ . For instance, the 5th partial sum of the Fourier series at hand can be explicitly described by adding in the previous cell the following:

```
FS = f.fourier_series_partial_sum(5, pi)
show("partial FS = ", FS)
```

To plot the suggested partial sums, together with  $f$ , you may add the block

```
FS5 = f.fourier_series_partial_sum(5, pi)
FS10 = f.fourier_series_partial_sum(10, pi)
FS20 = f.fourier_series_partial_sum(20, pi)
FS100 = f.fourier_series_partial_sum(100, pi)
a = f.plot(x, -pi, pi)
b5 = plot(FS5, x, -pi, pi, linestyle="--", tick_formatter=[2*pi, None])
b10 = plot(FS10, x, -pi, pi, linestyle="--", tick_formatter=[2*pi, None])
b20 = plot(FS20, x, -pi, pi, linestyle="--", tick_formatter=[2*pi, None])
b100 = plot(FS100, x, -pi, pi, linestyle="--", tick_formatter=[2*pi, None])
(a+b5).show(); (a+b10).show()
(a+b20).show(); (a+b100).show()
```

This procedure yields the figures presented below, illustrating the improvement in approximation as the number of terms is increased.

functions in a reasonable way. Nevertheless, the first statement of the theorem says that the best approximation possible (in the  $L_2$ -distance), is by the partial sums as described.

The second and third statements can be perceived as an analogy to the orthogonal projections onto subspaces in terms of Cartesian coordinates. Indeed, if for a given function  $g$ , the series  $F(x) = \sum_{n=1}^\infty c_n f_n(x)$  converges pointwise, then the function  $F(x)$  is, in a certain sense, the orthogonal projection of  $g$  into the vector subspace of all such series.

The second statement is called *Bessel's inequality* and it is an analogy of the finite-dimensional proposition that the size of the orthogonal projection of a vector cannot be larger than the original vector. The equality from the third statement is called *Parseval's theorem* and it says that if a given vector does not decrease in length by the orthogonal projection onto a given subspace, then it belongs to this subspace.

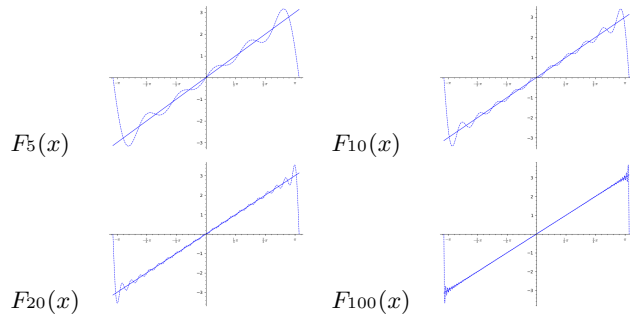
On the other hand, the theorem does not claim that the partial sums of the considered series need to converge pointwise to some function. There is no analogy to this phenomenon in the finite-dimensional world. In general, the series  $F(x)$  need not be convergent for some points  $x$ , even under the assumption of the equality in (3). However, if the series  $\sum_{n=1}^\infty |c_n|$  converges to a finite value, and if all the functions  $f_n$  are bounded uniformly on  $I$ , then, the series  $F(x) = \sum_{n=1}^\infty c_n f_n(x)$  converges at every point  $x$ . Yet it need not converge to the function  $g$  everywhere. We return to this problem later.

The proof of all of the three statements of the theorem is similar to the case of finite-dimensional Euclidean spaces. That is to be expected since the bounds for the distances of  $g$  from the partial sum  $f$  are constructed in the finite-dimensional linear hull of the functions concerned:

**PROOF OF THEOREM 7.1.5.** Choose any linear combination  $f = \sum_{n=1}^k a_n f_n$  and calculate its distance from  $g$ . We obtain

$$\begin{aligned} \|g - \sum_{n=1}^k a_n f_n\|^2 &= \int_a^b \left| g(x) - \sum_{n=1}^k a_n f_n(x) \right|^2 dx \\ &= \int_a^b |g(x)|^2 dx - \int_a^b \sum_{n=1}^k g(x) \overline{a_n f_n(x)} dx - \\ &\quad - \int_a^b \sum_{n=1}^k a_n f_n(x) \overline{g(x)} dx + \int_a^b \left| \sum_{n=1}^k a_n f_n(x) \right|^2 dx \\ &= \|g\|^2 - \sum_{n=1}^k \overline{a_n} c_n \|f_n\|^2 - \sum_{n=1}^k a_n \overline{c_n} \|f_n\|^2 + \sum_{n=1}^k a_n^2 \|f_n\|^2 \\ &= \|g\|^2 + \sum_{n=1}^k \|f_n\|^2 (|c_n - a_n|^2 - |c_n|^2). \end{aligned}$$

Since we are free to choose  $a_n$  as we please, we minimize the last expression by choosing  $a_n = c_n$ , for each  $n$ . This



□

**7.A.11. Triangle wave.** Find the Fourier series  $F(x)$  for the  $2\pi$ -periodic extension of the function  $f(x) = |x|$ , with  $x \in (-\pi, \pi)$ . Next present the 1st, 3rd and 6th partial sum of the Fourier series and use Sage to illustrate them for  $x \in (-4\pi, 4\pi)$ .

**Solution.** The given function corresponds to a sawtooth-shaped oscillation. Its expression as a Fourier series is very important in practice. The function  $g$  is even on  $(-\pi, \pi)$ , i.e.,  $f(-x) = f(x)$  for all  $x \in (-\pi, \pi)$ . Therefore, we have  $b_n = 0$  for all  $n \in \mathbb{Z}_+$ , and it suffices to determine  $a_n$ , for  $n \in \mathbb{N}$ . Combining the definition of  $a_n$  with 6.B.35, we get

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) dx = \frac{2}{\pi} \int_0^{\pi} x dx = \frac{2}{\pi} \left[ \frac{x^2}{2} \right]_0^{\pi} = \pi.$$

To compute  $a_n$  for  $n \in \mathbb{Z}_+$ , one can apply integration by parts:

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx = \frac{2}{\pi} \int_0^{\pi} x \cos(nx) dx \\ &= \frac{2}{\pi} \left[ \frac{x}{n} \sin(nx) \right]_0^{\pi} - \frac{2}{n\pi} \int_0^{\pi} \sin(nx) dx \\ &= \frac{2}{n^2\pi} [\cos(nx)]_0^{\pi} = \frac{2}{n^2\pi} ((-1)^n - 1). \end{aligned}$$

So

$$a_n = \begin{cases} -\frac{4}{n^2\pi}, & \text{for } n \text{ odd;} \\ 0, & \text{for } n \text{ even,} \end{cases}$$

and the Fourier series in question has the form

$$\begin{aligned} F(x) &= \frac{\pi}{2} + \frac{2}{\pi} \sum_{n=1}^{\infty} \left( \frac{(-1)^n - 1}{n^2} \cos(nx) \right) \\ &= \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2} \\ &= \frac{\pi}{2} - \frac{4}{\pi} \left( \cos x + \frac{\cos(3x)}{3^2} + \frac{\cos(5x)}{5^2} + \dots \right). \end{aligned}$$

completes the proof of the first statement. With this choice of  $a_n$ , we obtain *Bessel's identity*

$$\|g - \sum_{n=1}^k c_n f_n\|^2 = \|g\|^2 - \sum_{n=1}^k |c_n|^2 \|f_n\|^2.$$

Since the left-hand side is non-negative, it follows that:

$$\sum_{n=1}^k |c_n|^2 \|f_n\|^2 \leq \|g\|^2.$$

Let  $k \rightarrow \infty$ . Since every non-decreasing sequence of real numbers which is bounded from above has a limit, it follows that

$$\sum_{n=1}^{\infty} c_n^2 \|f_n\|^2 \leq \|g\|^2,$$

which is Bessel's inequality.

If equality occurs in Bessel's inequality, then statement (3) follows straight from the definitions and the Bessel's identity proved above. □

An orthogonal system of functions is called a *complete orthogonal system* on an interval  $I = [a, b]$  for some space of functions on  $I$  if and only if Parseval's equality holds for every function  $g$  in this space.

**7.1.6. Fourier series.** The coefficients  $c_n$  from the previous theorem are called the *Fourier coefficients* of a given function in the (abstract) *Fourier series*.



The previous theorem indicates that we are able to work with countable orthogonal systems of functions  $f_n$  in much the same way as with finite orthogonal bases of vector spaces.

There are, however, essential differences:

- It is not easy to decide what the set of convergent or uniformly convergent series

$$F(x) = \sum_{n=1}^{\infty} c_n f_n(x)$$

looks like.

- For a given integrable function, we can find only the "best approximation possible" by such a series  $F(x)$  in the sense of  $L_2$ -distance.

In the case when we have an orthonormal system of functions  $f_n$ , the formulae mentioned in the theorem are simpler, but still there is no further improvement in the approximations.

The choice of an orthogonal system of functions for use in practice should address the purpose for which the approximations are needed. The name "Fourier series" itself refers to the following choice of a system of real-valued functions:

FOURIER'S ORTHOGONAL SYSTEM

The following functions form an orthogonal system

$$1, \sin x, \cos x, \sin 2x, \cos 2x, \dots, \sin nx, \cos nx, \dots$$

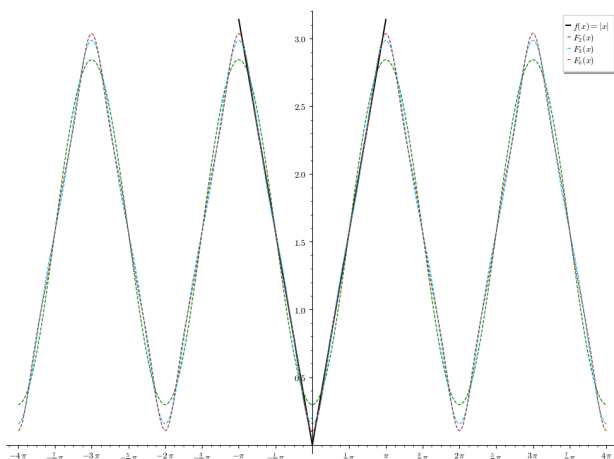
Thus

$$F_1(x) = \frac{\pi}{2} - \frac{4 \cos(x)}{\pi},$$

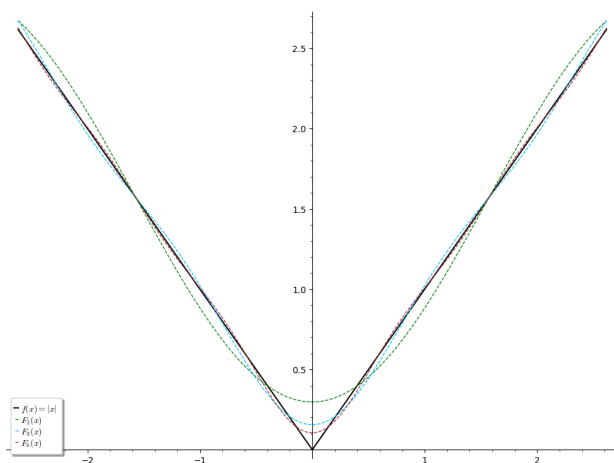
$$F_3(x) = F_1(x) - \frac{4 \cos(3x)}{9\pi},$$

$$F_6(x) = F_3(x) - \frac{4 \cos(5x)}{25\pi}.$$

Below we illustrate  $F_1, F_3$  and  $F_6$  together in a single figure:



To better observe the convergence of the Fourier series in a small neighbourhood around the origin, we may zoom in on the graph at that point:



Notice the Fourier series described here could be derived by easier means, namely by integrating the Fourier series of the square wave function (Heaviside's function), see 7.1.9. □

**7.A.12.** Use Sage to provide a solution of the task in 7.A.11. Moreover, show that  $F_5(x) = F_6(x)$  for all  $x$ . ○

**7.A.13.** Compute the Fourier series of the  $2\pi$ -periodic extension of the sign function  $\text{sgn} : [-\pi, \pi] \rightarrow \mathbb{R}$ , where recall that

$$\text{sgn}(x) = \begin{cases} -1, & \text{for } -\pi \leq x < 0, \\ 0, & \text{for } x = 0, \\ 1, & \text{for } 0 < x \leq \pi. \end{cases} \quad \circ$$

An elementary exercise on integration by parts shows that this is an orthogonal system of functions on the interval  $[-\pi, \pi]$ .

These functions are periodic with common period  $2\pi$  (see the definition below). “Fourier analysis”, which builds upon this orthogonal system, allows us to work with all piecewise continuous periodic functions with extraordinary efficiency. Since many physical, chemical, and biological data are perceived, received, or measured, in fact, by frequencies of the signals (the measured quantities), it is really an essential mathematical tool. Biologists and engineers often use the word “signal” in the sense of “function”.

PERIODIC FUNCTIONS

A real or complex valued function  $f$  defined on  $\mathbb{R}$  is called a *periodic function* with period  $T > 0$  if  $f(x + T) = f(x)$  for every  $x \in \mathbb{R}$ .

It is evident that sums and products of periodic functions with the same period are again periodic functions with the same period.

We note that the integral  $\int_{x_0}^{x_0+T} f(x) dx$  of a periodic function  $f$  on an interval whose length equals the period  $T$  is independent of the choice of  $x_0 \in \mathbb{R}$ . To prove it, it is enough to suppose  $0 \leq x_0 < T$ , using a translation by a suitable multiple of  $T$ . Then,

$$\begin{aligned} \int_{x_0}^{x_0+T} f(x) dx &= \int_{x_0}^T f(x) dx + \int_T^{x_0+T} f(x) dx \\ &= \int_{x_0}^T f(x) dx + \int_0^{x_0} f(x) dx = \int_0^T f(x) dx. \end{aligned}$$

Next, remind the Fourier's orthogonal system with the norms over intervals of length  $2\pi$  equal to

$$\|\sin nx\|^2 = \|\cos nx\|^2 = \pi$$

(cf. 6.2.6). Using it to built series as in the theorem 7.1.5, we arrive at

FOURIER SERIES

The series of functions

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx))$$

with coefficients

$$a_n = \frac{1}{\pi} \int_{x_0}^{x_0+2\pi} g(x) \cos(nx) dx,$$

$$b_n = \frac{1}{\pi} \int_{x_0}^{x_0+2\pi} g(x) \sin(nx) dx,$$

is called the *Fourier series* of a function  $g$  on the interval  $[x_0, x_0 + 2\pi]$ .

The coefficients  $a_n$  and  $b_n$  are called *Fourier coefficients of the function*  $g$ .

**7.A.14.** Use Sage to describe the Fourier series of the  $2\pi$ -periodic extension of the function

$$f(x) = \begin{cases} 0, & x \in [-\pi, 0), \\ \sin(x), & x \in [0, \pi). \end{cases}$$

Additionally, plot the graphs of  $F_3, F_5, F_8$  and  $f$  in a single figure, and specify the 8th and 48th coefficients of the corresponding Fourier series.

**Solution.** We have  $T = 2\pi$  and we are interested in the interval  $[-\pi, \pi)$ . Let us begin by computing the Fourier coefficients of  $f$ , which can be done with the method applied above, involving the commands `piecewise` and `integral`:

```
f1(x)=0; f2(x)=sin(x)
f=piecewise([( -pi, 0), f1], [(0, pi), f2])
L=pi; n=var("n")
an=(1/L)*integral(f1(x)*cos(n*pi*x/L), x, -pi, 0)
+(1/L)*integral(f2(x)*cos(n*pi*x/L), x, 0, pi)
show("a_n=", an.expand())
bn=(1/L)*integral(f1(x)*sin(n*pi*x/L), x, -pi, 0)
+(1/L)*integral(f2(x)*sin(n*pi*x/L), x, 0, pi)
show("b_n=", bn.expand())
```

Running this, we get the answers

$$a_n = -\frac{\cos(\pi n)}{\pi(n^2 - 1)} - \frac{1}{\pi(n^2 - 1)}, \quad b_n = -\frac{\sin(\pi n)}{\pi(n^2 - 1)},$$

which both make sense for all  $n \neq 1$ . Thus, for all positive integers  $n \neq 1$  we have

$$a_n = -\frac{1}{\pi(n^2 - 1)}((-1)^n + 1), \quad b_n = 0$$

and moreover  $a_0 = \frac{2}{\pi}$ . For  $n = 1$  we have the integrals

$$a_1 = \frac{1}{2\pi} \int_0^\pi \sin(2x) dx, \quad b_1 = \frac{1}{\pi} \int_{-\pi}^\pi f(x) \sin(x) dx,$$

and it is easy to see that  $a_1 = 0$  and  $b_1 = 1/2$ . In this way we arrive at the Fourier series

$$F(x) = \frac{1}{\pi} + \frac{\sin x}{2} + \frac{1}{\pi} \sum_{n=2}^\infty \left( \frac{(-1)^n + 1}{1 - n^2} \cos(nx) \right).$$

Since  $(-1)^n + 1 = 0$  when  $n$  is odd, and  $(-1)^n + 1 = 2$  when  $n$  is even, put  $n = 2m$  to obtain the expression

$$F(x) = \frac{1}{\pi} + \frac{\sin x}{2} - \frac{2}{\pi} \sum_{m=1}^\infty \frac{\cos(2mx)}{4m^2 - 1}.$$

For the partial sums we see that

$$F_3(x) = \frac{1}{\pi} + \frac{\sin(x)}{2} - \frac{2 \cos(2x)}{3\pi},$$

$$F_5(x) = -\frac{2 \cos(4x)}{15\pi} + F_3(x),$$

$$F_8(x) = -\frac{2 \cos(8x)}{63\pi} - \frac{2 \cos(6x)}{35\pi} + F_5(x).$$

In Sage to obtain these expressions we may continue typing in the preceding block the syntax given here:



If  $T$  is the time taken for one revolution of an object moving round the unit circle at constant speed, then that constant speed is  $\omega = 2\pi/T$ . In practice, we often want to work with Fourier series with an arbitrary primary period  $T$  of the functions, not just  $2\pi$ . Then we should employ the functions  $\cos(\omega nx)$ ,  $\sin(\omega nx)$ , where  $\omega = \frac{2\pi}{T}$ . By substitution  $t = \omega x$ , we can verify the orthogonality of the new system of functions and recalculate the coefficients in the Fourier series  $F(x)$  of a function  $g$  on the interval  $[x_0, x_0 + T]$ :

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^\infty (a_n \cos(n\omega x) + b_n \sin(n\omega x)),$$

which have values

$$a_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \cos(n\omega x) dx,$$

$$b_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \sin(n\omega x) dx.$$

**7.1.7. The complex Fourier coefficients.** Parametrize the unit circle in the form:

$$e^{i\omega t} = \cos \omega t + i \sin \omega t.$$

For all integers  $m, n$  with  $m \neq n$ ,

$$\begin{aligned} \int_{-\pi}^\pi e^{imx} e^{-inx} dx &= \int_{-\pi}^\pi e^{i(m-n)x} dx \\ &= \frac{1}{i(m-n)} [e^{i(m-n)x}]_{-\pi}^\pi = 0. \end{aligned}$$

Thus for  $m \neq n$ , the integral  $\langle e^{imx}, e^{inx} \rangle = 0$ .

FOURIER'S COMPLEX ORTHOGONAL SYSTEM

The following functions form a complex valued orthogonal system

$$e^{-in\omega t}, \dots, e^{-i\omega t}, 1, e^{i\omega t}, e^{i2\omega t}, \dots, e^{in\omega t}, \dots$$

Note that if  $m = n$ , then

$$\int_{-\pi}^\pi e^{imx} e^{-imx} dx = \int_{-\pi}^\pi dx = 2\pi.$$

The orthogonality of this system can be easily used to recover the orthogonality of the real Fourier's system: Rewrite the above result as

$$\int_{-\pi}^\pi (\cos mx + i \sin mx)(\cos nx - i \sin nx) dx = 0$$

By expanding and separating into real and imaginary parts we get both

$$\int_{-\pi}^\pi (\cos mx \cos nx + \sin mx \sin nx) dx = 0$$

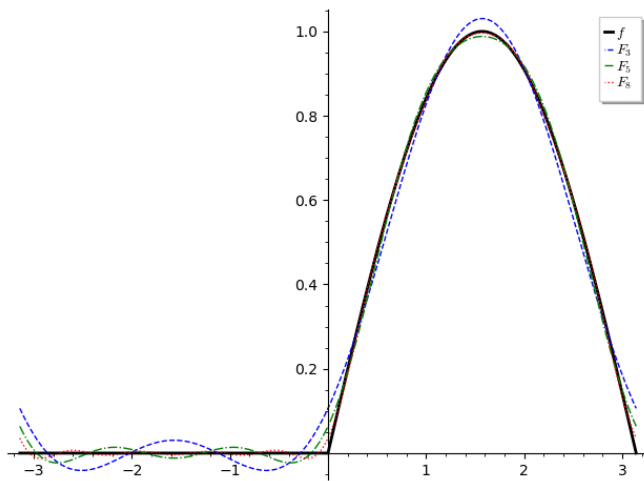
$$\int_{-\pi}^\pi (\sin mx \cos nx - \cos mx \sin nx) dx = 0$$

```

FS3 = f.fourier_series_partial_sum(3,pi)
FS5 = f.fourier_series_partial_sum(5,pi)
FS8 = f.fourier_series_partial_sum(8,pi)
show("3rd partial sum = ",FS3)
show("5th partial sum = ",FS5)
show("8th partial sum = ",FS8)
a = f.plot(x, -pi, pi, thickness=2,
color="black", legend_label="$f$")
b = plot(FS3, x,-pi,pi,linestyle="--",
color="blue",legend_label="$F_3$")
c= plot(FS5, x,-pi,pi,linestyle="-.",
color="green", legend_label="$F_5$")
d= plot(FS8, x, -pi, pi, linestyle=":",
color="red", legend_label="$F_8$")
(a+b+c+d).show()

```

This block also constructs the required graphs, which we present below.



Finally, for calculating the Fourier coefficients Sage provides built-in functions. For instance, one can use the syntax `f.fourier_series_cosine_coefficient(k)`, which corresponds to the  $k$ th coefficient of the cosine Fourier series of our periodic function  $f$ . Notice that if you are interested in “sine coefficients”, then the right function is `f.fourier_series_sine_coefficient(k)`, and in this case one should ensure that uses the sine function correctly, see 7.D.8 for an example.

Now you may check yourselves the desired coefficients, by simply adding the following cell:

```

Fc8=f.fourier_series_cosine_coefficient(8)
show("8th coefficient = ",Fc8)
Fc48=f.fourier_series_cosine_coefficient(48)
show("48th coefficient = ",Fc48)

```

In 7.D.11 we will present an alternative method using the Sage functions introduced in this block to determine and illustrate certain partial sums of a Fourier series, without using the Sage function `fourier_series_partial_sum`. □

By replacing  $n$  with  $-n$ , we have also

$$\int_{-\pi}^{\pi} (\cos mx \cos nx - \sin mx \sin nx) dx = 0$$

$$\int_{-\pi}^{\pi} (\sin mx \cos nx + \cos mx \sin nx) dx = 0$$

and hence, with  $m \neq n$ ,

$$\int_{-\pi}^{\pi} \cos mx \cos nx dx = 0$$

$$\int_{-\pi}^{\pi} \sin mx \sin nx dx = 0$$

$$\int_{-\pi}^{\pi} \sin mx \cos nx dx = 0$$

which proves again the orthogonality of the real valued Fourier system.

Note the case  $m = n > 0$ , when we recovered again

$$\int_{-\pi}^{\pi} \cos^2 nx dx = \|\cos(nx)\|^2 = \pi,$$

$$\int_{-\pi}^{\pi} \sin^2 nx dx = \|\sin(nx)\|^2 = \pi.$$

If  $n = 0$ , then  $\|1\|^2 = 2\pi$ .

For a (real or complex) function  $f(t)$  with  $-T/2 \leq t \leq T/2$ , and all integers  $n$ , we can define, in this context, its *complex Fourier coefficients* by the complex numbers

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-i\omega n t} dt.$$

The relation between the coefficients  $a_n$  and  $b_n$  of the Fourier series (after recalculating the formulae for these coefficients for functions with a general period of length  $T$ ) and these complex coefficients  $c_n$  follow from the definitions. Clearly,  $c_0 = a_0/2$ , and for natural numbers  $n$ , we have

$$c_n = \frac{1}{2}(a_n - ib_n), \quad c_{-n} = \frac{1}{2}(a_n + ib_n).$$

If the function  $f$  is real valued,  $c_n$  and  $c_{-n}$  are complex conjugates of each other.

We note here that for real valued functions with period  $2\pi$ , the Bessel inequality in this notation becomes

$$\frac{1}{2}|a_0|^2 + \sum_{n=1}^{\infty} (|a_n|^2 + |b_n|^2) \leq \|f\|^2 = \frac{1}{\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt.$$

We have expressed the Fourier series  $F(t)$  for a function  $f(t)$  in the form

$$F(t) = \sum_{n=-\infty}^{\infty} c_n e^{i\omega n t}.$$

In both cases of real or complex valued functions, the corresponding Fourier series can be written in this form. In general, the coefficients are complex. We return to this expression later, in particular when dealing with Fourier transforms.

For fixed  $T$ , the expression  $\omega = 2\pi/T$  describes how the frequency changes if  $n$  is increased by one.

**7.A.15. Fourier series over other intervals.** Consider the function  $g : [-1, 1) \rightarrow \mathbb{R}$ , defined by

$$g(x) = \begin{cases} 0, & \text{for } -1 \leq x < 0, \\ x + 1, & \text{for } 0 \leq x < 1. \end{cases}$$

- (a) Determine a 2-periodic extension  $\hat{g}$  of  $g$  for  $x \in [-3, 3)$  and next sketch its graph via Sage.
- (b) Find the corresponding Fourier series on  $[-1, 1)$ , and next confirm your result via Sage.
- (c) Use Sage to plot the partial sums  $F_5(x)$ , and  $F_{55}(x)$  together with  $\hat{g}$ , for  $x \in (-3, 3)$ .

**Solution.** (a) By assumption the period is  $T = 2$ , and on the interval  $[-3, 3)$  a 2-periodic extension of  $g$  has the form

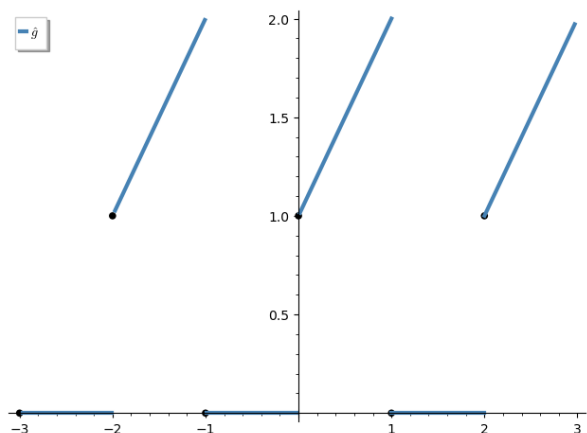
$$\hat{g}(x) = \begin{cases} g(x+2), & \text{for } -3 \leq x < -1, \\ g(x), & \text{for } -1 \leq x < 1, \\ g(x-2), & \text{for } 1 \leq x < 3. \end{cases}$$

$$= \begin{cases} 0, & \text{for } -3 \leq x < -2, \\ x+3, & \text{for } -2 \leq x < -1, \\ 0, & \text{for } -1 \leq x < 0, \\ x+1, & \text{for } 0 \leq x < 1, \\ 0, & \text{for } 1 \leq x < 2, \\ x-1, & \text{for } 2 \leq x < 3. \end{cases}$$

In Sage we can use the `piecewise` command to plot  $\hat{g}$ , so one may try the following block:

```
g1(x)=0;g2(x)=x+1;g3(x)=x+3;g4(x)=x-1
g=piecewise([( (-3,-2),g1), [ (-2,-1),g3],
[ (-1,0),g1], [ (0,1),g2], [ (1,2),g1], [ (2,3),g4] ])
a = g.plot(x, -3, 3,thickness=3,
color="steelblue",
exclude=[-2, -1, 0, 1, 2],
legend_label=r"\hat{g}")
p0=point([-3, 0], size=30, color="black")
p1=point([-2, 1], size=30, color="black")
p2=point([-1, 0], size=30, color="black")
p3=point([0, 1], size=30, color="black")
p4=point([1, 0], size=30, color="black")
p5=point([2, 1], size=30, color="black")
show(a+p0+p1+p2+p3+p4+p5)
```

Let us present the result:



We wish to show that the Fourier orthogonal system is complete on  $\mathcal{S}^0[-\pi, \pi]$ . This needs a thorough technical preparation. So now, we only formulate the main result, and add some practical notes. The proof of the theorem will be discussed later, see 7.1.15.

**7.1.8. Theorem.** Consider a bounded interval  $[a, b]$  of length  $T = b - a$ . Let  $f$  be a real or complex valued function in  $\mathcal{S}^1[a, b]$  (i.e., a piecewise continuous function with a piecewise continuous first derivative), extended periodically on  $\mathbb{R}$ . Then:

- (1) The partial sums  $s_N$  of its Fourier series converge pointwise to the function

$$g(x) = \frac{1}{2} \left( \lim_{y \rightarrow x+} f(y) + \lim_{y \rightarrow x-} f(y) \right).$$

- (2) Moreover, if the periodic extension of  $f \in \mathcal{S}^1[a, b]$  is a continuous function on  $\mathbb{R}$ , then the pointwise convergence of its Fourier series is uniform.
- (3) The  $L_2$ -distance  $\|s_N - f\|$  converges to 0 for  $N \rightarrow \infty$ .

**7.1.9. Periodic extension of functions.** The Fourier series converges, of course, outside the original interval  $[-T/2, T/2]$ , since it is a periodic function on  $\mathbb{R}$ . The Heaviside function  $g$  is defined by



$$g(x) = \begin{cases} -1 & \text{if } -\pi < x < 0, \\ 1 & \text{if } 0 < x < \pi \end{cases}$$

(We do not need to define the values at zero and at the end points of the interval, because these do not effect the coefficients of the Fourier series.) The periodic extension of the Heaviside function onto all of  $\mathbb{R}$  is usually called the *square wave function*.

Since  $g$  is an odd function, the coefficients of the functions  $\cos(nx)$  are all zero. The coefficients of  $\sin(nx)$ , are given by

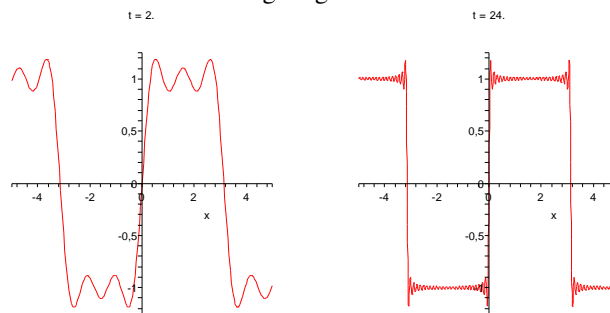
$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(x) \sin(nx) dx = \frac{2}{\pi} \int_0^{\pi} \sin(nx) dx$$

$$= \frac{2}{n\pi} (1 - (-1)^n).$$

Thus the Fourier series of  $g$  is

$$g(x) = \frac{4}{\pi} \left( \sin(x) + \frac{1}{3} \sin(3x) + \frac{1}{5} \sin(5x) + \dots \right).$$

The partial sums of its first five and fifty terms, respectively, are shown in the following diagrams.



Notice that we have included the points where the jump discontinuities occur within the interval  $[-3, 3)$ , using the command point.

(b) In this task one should apply the general formulae given in 7.1.6. We have  $\omega = 2\pi/T = \pi$ , and

$$a_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \cos(n\omega x) dx, \forall n \in \mathbb{N},$$

$$b_n = \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \sin(n\omega x) dx, \forall n \in \mathbb{Z}_+.$$

Now it is easy to see that  $a_0 = \int_{-1}^1 g(x) dx = \int_0^1 (x+1) dx = \frac{3}{2}$  and moreover

$$a_n = \int_{-1}^1 g(x) \cos(n\pi x) dx = \int_0^1 (x+1) \cos(n\pi x) dx$$

$$= \frac{(-1)^n - 1}{n^2\pi^2},$$

$$b_n = \int_{-1}^1 g(x) \sin(n\pi x) dx = \int_0^1 (x+1) \sin(n\pi x) dx$$

$$= \frac{1 - 2(-1)^n}{n\pi}.$$

In Sage to confirm these computations, it suffices to use only  $g$  (and not its extension), as follows:

```
g1(x)=0;g2(x)=x+1
g=piecewise([( -1, 0), g1], [(0, 1), g2])
T=2; n=var("n"); Om=pi
an=(2/T)*integral(g1(x)*cos(n*Om*x), x, -1, 0) \
+(2/T)*integral(g2(x)*cos(n*Om*x), x, 0, 1)
show("a_n=", an.expand())
bn=(2/T)*integral(g1(x)*sin(n*Om*x), x, -1, 0) \
+(2/T)*integral(g2(x)*sin(n*Om*x), x, 0, 1)
show("b_n=", bn.expand()); integral(x+1, x, 0, 1)
```

In this way we obtain the following expression of the Fourier series in question:

$$\frac{3}{4} + \sum_{n=1}^{\infty} \left( \frac{(-1)^n - 1}{n^2\pi^2} \cos(n\pi x) + \frac{1 - 2(-1)^n}{n\pi} \sin(n\pi x) \right),$$

and one can try to obtain refinements by considering odd/even cases.

(c) This task can be solved by the following block:

If the interval  $[-T/2, T/2]$  is chosen for the prime period  $T$  of such a square wave function, the resulting Fourier series is

$$g(x) = \frac{4}{\pi} \left( \sin(\omega x) + \frac{1}{3} \sin(3\omega x) + \frac{1}{5} \sin(5\omega x) + \dots \right).$$

The number  $\omega = \frac{2\pi}{T}$  is also called the *phase frequency* of the wave.

As the number of terms of the series increases, the approximation gets much better except in a (still shrinking) neighbourhood of the discontinuity point. There, the maximum of the deviation remains roughly the same. This is a general property of Fourier series and it is called the *Gibbs phenomenon*.

In accordance with 7.1.8(1), the Fourier series of the Heaviside function converges to the mean of the two one-sided limits at the points of discontinuity.

Of course, we cannot expect that the convergence of Fourier series for functions  $g$  with discontinuity points be uniform (then, the function  $g$  would have to be continuous itself, being a uniform limit of continuous functions).

**7.1.10. Utilizing symmetry of functions.** We consider the



problem of finding the Fourier series of the function  $f(x) = x^2$  on the interval  $[0, 1]$ . If we just periodically extend this function from the given interval  $[0, 1]$ , the resulting function would not be continuous, and so the convergence at integers would be as slow as in the case of a square wave function. However, we can easily extend the domain of  $f$  to the interval  $[-1, 1]$ , so that  $f(x) = x^2$  is an even function for  $-1 \leq x \leq 1$ . If we then extend periodically, the result is continuous. The resulting Fourier series then converges uniformly, and since then  $f$  is even, only the coefficients  $a_n$  are non-zero.

For  $n > 0$ , iterated application of integration by parts yields:

$$a_n = \frac{2}{2} \int_{-1}^1 x^2 \cos\left(\frac{2\pi nx}{2}\right) dx = 2 \int_0^1 x^2 \cos(\pi nx) dx$$

$$= \frac{4}{\pi^2 n^2} (-1)^n.$$

The remaining coefficient is

$$a_0 = \frac{2}{2} \int_{-1}^1 x^2 dx = 2 \int_0^1 x^2 dx = \frac{2}{3}.$$

The entire series giving the periodic extension of  $x^2$  from the interval  $[-1, 1]$  thus equals

$$g(x) = \frac{1}{3} + \frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos(\pi nx).$$

By the Weierstrass criterion, this series converges uniformly. Therefore,  $g(x)$  is continuous. By theorem 7.1.8,  $g(x) = f(x) = x^2$  on the interval  $[-1, 1]$ . Thus our series approximates the function  $x^2$  on the interval  $[0, 1]$  better (i.e. faster) than we could achieve with the periodic extension of the function from  $[0, 1]$  interval only. If we put  $x = 1$  and rearrange,



```
g1(x)=0;g2(x)=x+1;g3(x)=x+3;g4(x)=x-1
g=piecewise([(−3, −2), g1], [(−2, −1), g3], [(0, 1), g2], [(1, 2), g1], [(2, 3), g4]])
partS5 = g.fourier_series_partial_sum(5,1)
partS55 = g.fourier_series_partial_sum(55,1)
a = g.plot(x, −3, 3, thickness=3, color="steelblue",
exclude=[−2, −1, 0, 1, 2], legend_label=r"$f$")
s5=plot(partS5, x, −3, 3, linestyle="--",
```

we obtain the remarkable result

$$\frac{1}{4}(g(1) - \frac{1}{3}) = \frac{\pi^2}{6} = \sum_{n=1}^{\infty} \frac{(-1)^{2n}}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

We proceed with a further illustration. Let us try to differentiate the Fourier series for  $x^2$  term by term and, of course we hope to get the Fourier series of  $x$  on the interval  $-1 < x < 1$

$$x = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(\pi nx).$$

If it was truly the series of  $x$ , it could not converge uniformly since the periodic extension of the function  $x$  from the interval  $[-1, 1]$  is not a continuous function. Thus our differentiation term by term is not guaranteed by the general theorem.

However, a straightforward computation of the Fourier coefficients shows, that this is the Fourier series of  $x$ .

The series for  $x^2$  converges uniformly, and so we can integrate it term by term obtaining

$$\frac{1}{3}x^3 = \frac{2}{3}x + \frac{4}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^3} \sin(\pi nx).$$

This is valid for  $-1 < x < 1$ . It is not valid for other values of  $x$ , since the series is periodic, but the other two terms are not. Of course, we may substitute the above Fourier series of the function  $x$  and thus obtain the Fourier series for  $x^3$  on the interval  $[-1, 1]$  this way. Notice, this will no more converge uniformly.

In this context, we use the following terminology for the Fourier series of even or odd functions:

THE SINE AND COSINE FOURIER SERIES

For a given (real or complex) valued function  $f$  on an interval  $[0, T)$  of length  $T > 0$ , the Fourier series of its even periodic extension (with period  $2T$ ) is called the *cosine Fourier series* of  $f$ , while the Fourier series of the odd periodic extension of  $f$  is called the *sine Fourier series* of the function  $f$ .

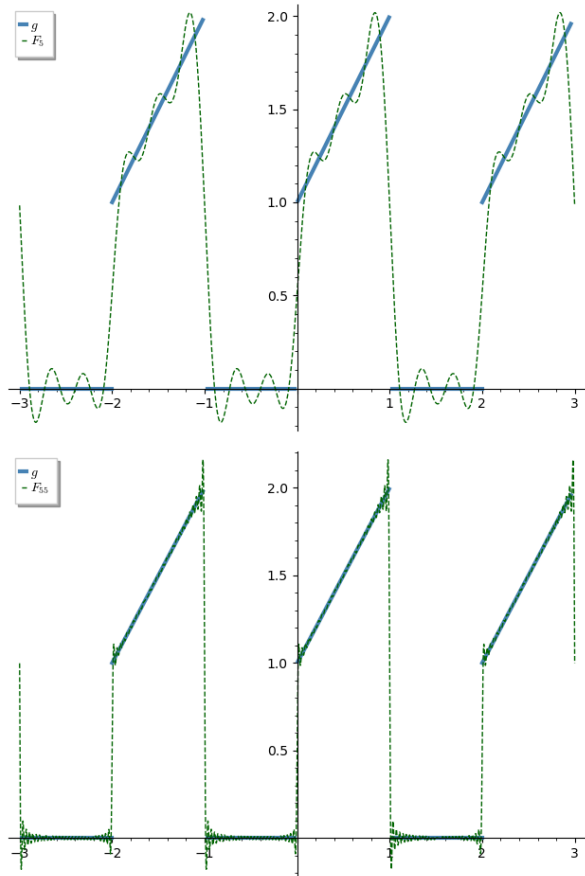
**7.1.11. General Fourier series and wavelets.** In the case of a general orthogonal system of functions  $f_n$  and the series generated from it, we often talk about the *general Fourier series* with respect to the orthogonal system of functions  $f_n$ .

Fourier series and further tools built upon them are used for processing various signals, images, and other data. In fact, these mathematical tools also underpin many fundamental models in science, including, for example, modeling of the function of the brain, as well as much of theoretical physics.

The periodic nature of the sine and cosine functions used in classical Fourier series, and their simple scaling by increasing the frequency by unit steps, limit their usability. In many

```
color="darkgreen", legend_label=r"$F_{5}$")
s55=plot(partS55, x, -3, 3, linestyle="--",
color="darkgreen", legend_label=r"$F_{55}$")
show(a+s5);show(a+s55)
```

Notice the resulting approximations are much slower comparing the situation in the previous examples. By the illustrations one can also demonstrate the “Gibbs phenomenon”. This is the overshooting in the jumps, which is proportional to the magnitudes of the jumps, see the figures here:



□

It is often convenient (and usually easier) to express the Fourier series using the complex coefficients  $c_n$  instead of the real coefficients  $a_n$  and  $b_n$ . This is a straightforward consequence of the facts

$$e^{in\omega x} = \cos(n\omega x) + i \sin(n\omega x) \quad \text{or, vice versa}$$

$$\cos(n\omega x) = \frac{1}{2}(e^{in\omega x} + e^{-in\omega x}),$$

$$\sin(n\omega x) = \frac{1}{2i}(e^{in\omega x} - e^{-in\omega x}).$$

The resulting series for a real or complex valued function  $g$  on the interval  $[-\pi, \pi]$  is  $F(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$  with

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-inx} g(x) dx.$$

applications, the nature of the data may suggest more convenient and possibly more efficient orthogonal systems of functions.

Requirements for fast numerical processing usually include quick scalability and the possibility of easy translations by constant values. In other words, we want to be able to zoom quickly in and out with respect to the frequencies and, at the same time, to localize in time.

Fast scalability can be achieved by having just one wavelet mother function<sup>1</sup>  $\psi$ , if possible with compact support, from which we create countably many functions  $\psi_{j,k}$ ,  $j, k \in \mathbb{Z}$ , by integer translations and dyadic dilations:

$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k).$$

It is wise to rescale and choose  $\psi$  with  $L_2$ -norm equal to one (as a function on  $\mathbb{R}$ ). Then the coefficients  $2^{j/2}$  ensure that the same is true for all  $\psi_{j,k}$ .

Of course, the shape of the mother wavelet  $\psi$  should match and cover all typical local behaviour of the data to be processed. We say that  $\psi$  is an orthogonal mother wavelet if the resulting countable system of functions  $\psi_{j,k}$  is orthogonal and, at the same time, it is “reasonably dense” in the space of functions with integrable squares. We come to this concept in more detail later on.

The effectivity of the wavelet analysis is another issue which needs further ideas and concepts to be built in. We do not have space here to go into details, but the readers may find many excellent specialized books in this fascinating area of applied Mathematics. Here we consider one simple example.

**7.1.12. The Haar wavelets.** Perhaps the first question to start with is, how to effectively approximate any given function with piecewise constant ones.



For various reasons, it is good if our mother wavelet  $\psi$  has zero mean, too. Thus we want to consider an analogue of the Heaviside function

$$\psi(x) = \begin{cases} 1 & x \in [0, 1/2) \\ -1 & x \in [1/2, 1). \end{cases}$$

As a straightforward exercise we may check that, indeed, the resulting system of functions  $\psi_{j,k}$  is orthonormal. Another exercise shows, using finite linear combinations of these functions, that we may approximate any constant function with given precision over a bounded interval. In an exercise we shall see ?? that this already verifies the density properties required for the orthogonal mother wavelet functions.

<sup>1</sup>The roots of wavelets go back to various attempts, how to localize the basic signals in both time and frequency, with diverse motivations from engineering and other applications. The name wavelet seems to be related to the idea of having a wave similar signal which begins and ends with zero amplitude. Since late 1970’s, these attempts were related to many names (e.g. Morlet, Meyer) and the wavelet theory became the main tool in signal analysis. Of course, first examples of wavelets are much older, the Haar’s construction goes back to 1909. Actually, many of the wavelet types do not represent orthogonal systems of functions, they rather share the idea of a combination of the high pass filters and low pass filters. The reader is advised to consult extremely rich literature, if interested in more details

Let us discuss a few examples and for more details see 7.1.7 (see also 7.D.10).

**7.A.16. Complex Fourier series.** (a) Determine the complex version of the Fourier series  $F(x)$  of the function  $f(x) = x$  with  $x \in (-\pi, \pi)$ . Next use the results from 7.A.10 to confirm that

$$c_n = \frac{1}{2}(a_n - ib_n), \quad c_{-n} = \overline{c_n}.$$

(b) Verify that the complex form of the Fourier series matches the result given in 7.A.10.

**Solution.** (a) By definition,  $F(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$ , hence one needs to specify the complex Fourier coefficients  $c_n$ . We have

$$c_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} x dx = 0,$$

since  $f(x) = x$  is an odd function. For  $n \neq 0$ , using integration by parts we get

$$\begin{aligned} c_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} x e^{-inx} dx \\ &= \frac{1}{2\pi} \left[ -\frac{x e^{-inx}}{in} \right]_{-\pi}^{\pi} - \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{-inx}}{-in} dx \\ &= -\frac{1}{2i\pi n} [x e^{-inx}]_{-\pi}^{\pi} + \frac{1}{2i\pi n} \int_{-\pi}^{\pi} e^{-inx} dx \\ &= -\frac{1}{2i\pi n} (\pi e^{-in\pi} + \pi e^{in\pi}) + \frac{1}{2i\pi n} \left[ -\frac{e^{-inx}}{in} \right]_{-\pi}^{\pi} \\ &= -\frac{2\pi \cos(n\pi)}{2\pi in} - \frac{1}{2\pi(in)^2} [e^{-inx}]_{-\pi}^{\pi} \\ &= -\frac{(-1)^n}{in} + \frac{1}{2\pi n^2} (e^{-in\pi} - e^{in\pi}) \\ &= -\frac{(-1)^n}{in} + \frac{-2i \sin(n\pi)}{2\pi n^2} = -\frac{(-1)^n}{in}. \end{aligned}$$

Here we used the identities  $i^2 = -1$ ,  $e^{-in\pi} + e^{in\pi} = 2 \cos(n\pi) = 2(-1)^n$  and  $e^{-in\pi} - e^{in\pi} = -2i \sin(n\pi) = 0$ .

Now, according to 7.A.10 we have  $a_n = 0$ ,  $b_n = \frac{2}{n}(-1)^{n+1}$ , and it should be

$$\begin{aligned} c_n &= \frac{1}{2}(a_n - ib_n) = \frac{1}{2} \left( 0 - i \frac{2}{n} (-1)^{n+1} \right) \\ &= \frac{-i(-1)^{n+1}}{n} = \frac{(-1)^{n+1}}{in}, \end{aligned}$$

since  $-i = \frac{1}{i}$ . This agrees with the result presented above and can be verified in Sage using the `bool` function. For example, you may type the cell

```
var("n")
bn=(2*(-1)^(n+1))/n; cn=(-(-1)^(n))/(i*n)
bool(cn==-(i*bn)/2)
```

which returns `True`. In a similar way, for  $c_{-n}$  one gets  $c_{-n} = \frac{(-1)^n}{in}$  and it is easy to see that  $c_n = \overline{c_{-n}}$ , that is,  $c_{-n}$  is the complex conjugate of  $c_n$ , see also 7.1.7.

Now we consider the question of effective treatment. Notice that we can also use the characteristic function  $\varphi$  of the interval  $[0, 1)$  and write

$$\psi(x) = \varphi(2x) - \varphi(2x-1) = \frac{1}{\sqrt{2}} \sqrt{2} \varphi_{1,0}(x) - \frac{1}{\sqrt{2}} \sqrt{2} \varphi_{1,1}(x).$$

The function  $\varphi$  plays the role of the *father wavelet function* and it itself satisfies

$$\varphi(x) = \varphi(2x) + \varphi(2x-1) = \frac{1}{\sqrt{2}} \sqrt{2} \varphi_{1,0}(x) + \frac{1}{\sqrt{2}} \sqrt{2} \varphi_{1,1}(x).$$

This can be interpreted as differencing and averaging the two consecutive values at the half scale.

With these properties, there is no need for an explicit analytic form of  $\psi$  and  $\varphi$ , since we can find their values recurrently in all dyadic points  $x$ . Indeed,

$$\varphi(2^{j-1}n) = \varphi(2^j n) + \varphi(2^j n - 1).$$

The function  $\varphi$  has another useful feature. Namely we can obtain the unit constant function by adding all its integer translations

$$\sum_{k=-\infty}^{\infty} \varphi_{0,k}(x) = 1$$

for all  $x \in \mathbb{R}$ .

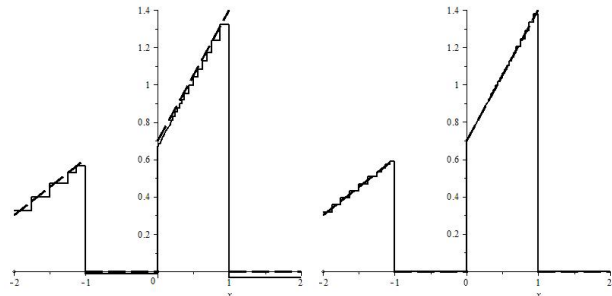
Finally, nearly all the coefficients in the general Fourier series with the base  $\psi_{j,k}$  vanish for piecewise constant functions. On the contrary, the function  $\varphi$  "sees" the constants. In engineering terminology, this is an instance of the *high pass filter* and *low pass filter*.

**7.1.13. Example.** To illustrate the above considerations, we approximate the following function  $f(x)$  in  $\mathbb{R}$  by the Haar wavelets,



$$f(x) = \begin{cases} 0.3(x+3), & -2 \leq x \leq -1 \\ 0.7(x+1), & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

The function in question is not periodic and could not be approximated well by classical Fourier series. The individual functions  $\psi_{j,k}$  from 7.1.11 have compact support, but in order to approximate constant or linear behaviour, we still need a large number of them. The following illustrations have been acquired in Maple working with indices  $|j| \leq n$  and  $|k| \leq n$ , the first one with  $n = 5$ , the second one with  $n = 10$ .



The approximation on the sides of the interval is not as good as in the middle, because we do not include enough shifts, i.e. values of  $k$ . One of the motivations for the scaling

For the complex expression of  $F(x)$  we finally get

$$\begin{aligned} F(x) &= c_0 + \sum_{n=1}^{\infty} c_n e^{inx} + \sum_{n=1}^{\infty} c_{-n} e^{-inx} \\ &= 0 + \sum_{n=1}^{\infty} \frac{-(-1)^n}{in} e^{inx} + \sum_{n=1}^{\infty} \frac{(-1)^n}{in} e^{-inx} \\ &= \sum_{n=1}^{\infty} \frac{(-1)^n}{in} (e^{-inx} - e^{inx}). \quad (b) \end{aligned}$$

(b) This follows directly from (b), combined with the identity  $-2i \sin(nx) = e^{-inx} - e^{inx}$ , i.e.,

$$\begin{aligned} F(x) &= \sum_{n=1}^{\infty} \frac{(-1)^n}{in} (e^{-inx} - e^{inx}) \\ &= \frac{1}{i} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} (-2i \sin(nx)) \\ &= \frac{i}{i} \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx), \end{aligned}$$

which is the formula presented in 7.A.10. □

**7.A.17.** Compute the complex Fourier series of the  $2\pi$ -periodic extension of the exponential map  $f(x) = e^x$ , with  $x \in (-\pi, \pi)$ . Next use Sage to illustrate the 20th-partial sum of the Fourier series together with that of  $f$ . ○

When  $f$  is a periodic function with period  $2L$  for some  $L > 0$ , then the complex Fourier series for  $x \in (x_0, x_0 + 2L)$  has the form  $F(x) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{i\pi n x}{L}}$ , where

$$c_n = \frac{1}{2L} \int_{x_0}^{x_0+2L} f(x) e^{-\frac{i\pi n x}{L}} dx.$$

The next task focuses on this case and is left as a challenge for you.

**7.A.18.** For the function  $f(x) = e^{-x}$  with  $x \in [-1, 1]$  and  $f(x) = f(x + 2)$  outside the interval  $[-1, 1]$ , compute the complex Fourier coefficients and derive the corresponding Fourier series. ○

Let  $f : [a, b] \rightarrow \mathbb{R}$  be a function for which  $f'$  exists. We say that  $f$  is piecewise differentiable if both  $f$  and  $f'$  are piecewise continuous on  $[a, b]$ . Suppose now that  $f$  is a  $2L$ -periodic piecewise differentiable function defined on an interval of length  $L$ , for some  $L > 0$ , and let  $F$  be the Fourier series of  $f$ . Assuming that  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ , the first part of the theorem in 7.1.8 (known as the “Dirichlet condition”, see 7.1.16) implies that at a point of continuity of  $f$ , the series  $F$  converges to the value of  $f$  at that point. Moreover, at a point of discontinuity the series  $F$  converges to the average of the left and right limits of  $f$  at that point. Thus, at a jump discontinuity  $x_0$  of  $f$ , the value of the Fourier series  $F$  does not depend on the value  $f(x_0)$ , but only on the left and right limits of  $f$  at  $x_0$ . This means that the behaviour of the Fourier series of  $f$  at discontinuity points is not effected

and shifting in the construction of wavelets is the hope for a small amount of non-zero coefficients. But this does not mean that most of the coefficients would be zero. In our case the percentage of non-zero coefficients for  $n = \{1, 2, \dots, 10\}$  is 55.6, 44., 38.8, 34.6, 32.2, 30.8, 29.8, 28.7, 28.0, 27.4.

**7.1.14. Concluding remarks.** A series of famous wavelets  $D_N$  has been called after Ingrid Debauchies. They are constructed by very similar recurrent averaging and differencing relations based on certain natural requirements. Just as an indication, consider the slightly more general recurrent relations

$$\begin{aligned} \varphi(x) &= \sqrt{2} \sum_{k=0}^N h_k \varphi(2x - k) \\ \psi(x) &= \sqrt{2} \sum_{k=0}^N g_k \varphi(2x - k) \end{aligned}$$

with yet unknown constants  $g_k$  and  $h_k$ . If we want the mother wavelet  $\psi(x-k)$  to have zero coefficients in the resulting series for all polynomials up to the order  $N-1$ , then we must ensure that

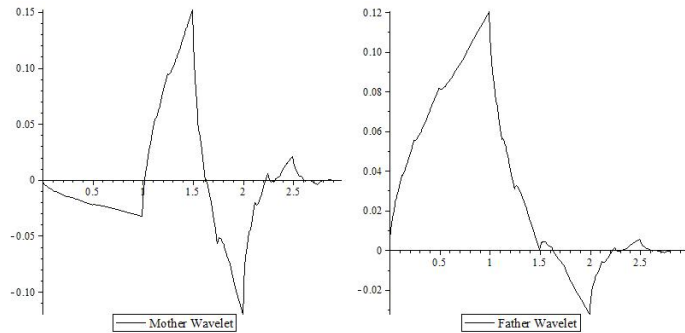
$$\int_{-\infty}^{\infty} x^k \psi(x) dx = 0$$

for all  $k = 0, 1, \dots, N-1$ . Similar conditions determine the Debauchies wavelets.

The standards of JPEG2000 are based on similar wavelets and such techniques provide tools for professional compression of visual data in film industry, or the format DjVu for compressed publications.

In the diagram below, there are the Daubechies  $D4$  mother and father wavelets.

In real applications, the orthogonality of the mother wavelet can be relaxed. As long as the functions  $\psi_{k,l}$  are linearly independent and generate the whole space of interest, we always get the dual basis with respect to the  $L_2$  inner product.



by the values of  $f$  there. On the other hand, recall that the continuity of  $f$  is a necessary and sufficient condition for the uniform convergence of  $F$  over the entire real line.

Let us describe a few examples and for further material on the convergence of Fourier series see the final section of this chapter (Section D).

**7.A.19.** Describe the implications of Dirichlet's condition for the Fourier series obtained in Problem 7.A.10.

**Solution.** Consider the function  $f(x) = x$ , with  $x \in (-\pi, \pi)$ , as in 7.A.10. The  $2\pi$ -periodic extension of  $f$  has discontinuities at  $x = \pm\pi, \pm3\pi, \pm5\pi, \dots$  and provides an example of a periodic piecewise differentiable function. At these discontinuities it is easy to see that the average of the left and right limits of  $f$  is zero. Therefore, by the result in 7.A.10 and Dirichlet's condition one deduces that

$$\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx) = \begin{cases} x, & \text{if } -\pi < x < \pi, \\ 0, & \text{if } x = \pm\pi. \end{cases}$$

□

**7.A.20. Convergence of Fourier series.** Consider the square wave function (Heaviside's function, see also 7.1.9), defined this time on  $[-1, 1]$ , i.e.,

$$f(x) = \begin{cases} -1, & \text{if } -1 \leq x < 0, \\ 1, & \text{if } 0 < x \leq 1, \end{cases}$$

with  $f(x+2) = f(x)$  for all  $x$  outside the interval  $[-1, 1]$ . Determine the pointwise convergence of the corresponding Fourier series.

**Solution.** To answer the task we do not need the explicit expression of the Fourier series  $F$  of  $f$ . In particular,  $f$  satisfies Dirichlet's condition and hence  $F$  converges to  $f$  at all points where  $f$  is continuous. The discontinuities of  $f$  appear at  $x = 0$ , and  $x = \pm 1$ . In particular, for the endpoints we have  $f(-1) \neq f(1)$  and the 2-periodic extension of  $f$  has discontinuities at  $\pm 1, \pm 3, \dots$ . At these points the Fourier series converges to the average of the left-hand and right-hand limits, which all equal to zero. For instance, we see that  $f(0^+) = \lim_{x \rightarrow 0^+} f(x) = 1$  and  $f(0^-) = \lim_{x \rightarrow 0^-} f(x) = -1$ , hence

$$\frac{1}{2}(f(0^-) + f(0^+)) = \frac{1}{2}(-1 + 1) = 0.$$

The limits  $f(0^-)$  and  $f(0^+)$  can be verified also in Sage, by utilizing the `piecewise` and `limit` commands. It is important however to specify the option `algorithm = "giac"` within the `limit` command, as shown below:

```
f = piecewise([[(-1, 0), -1], [(0, 1), 1]])
show(f(x).limit(x=0, dir="+", algorithm="giac"))
show(f(x).limit(x=0, dir="-", algorithm="giac"))
```

Thus  $F(x)$  converges to  $f(x)$  for all  $x$ , except of  $x = 0, \pm 1$  where it converges to 0. □

**7.1.15. Proof of theorem 7.1.8 about Fourier series.** We



return to the detailed proof of the basic properties of the classical Fourier series. We shall need several concepts and technicalities related to abstract metric spaces. Thus, the reader could enjoy reading first the general context of metrics and convergence which we will introduce in the beginning of the last part of this chapter. In this perspective, reading first a few paragraphs starting by 7.3.1, and returning later might be a good idea. On the other hand, we do not need much from the general theory of metric spaces and so our considerations of various concepts of convergence in the proof could also be of assistance for the abstract developments later.

We do not worry here about necessary conditions for convergence, and many other formulations can be found in literature. On the other hand, the statement of theorem 7.1.8 is quite simple and deals with many useful situations, as we have seen already.

Although we need only the  $L_1$  and  $L_2$  norms now, we should observe that for general  $1 \leq p < \infty$ , the formula

$$\|f\|_p = \left( \int_a^b |f|^p \right)^{1/p}$$

defines also a norm. See the definition in 7.3.1 and the paragraph on the  $L_p$ -norms and Hölder inequality in 7.3.4 below. Moreover, there is the  $L_\infty$  norm given by the suprema of values of  $f$  over the interval in question.

For the sake of simplicity, we always work in the space  $S_c^0$  or  $S_c^1$  with respect to the corresponding norm (which always makes sense there).

Hölder's inequality (applied to the functions  $f$  and constant 1) yields the following bound on  $S^0[a, b]$ . Namely, for  $p > 1$  and  $1/p + 1/q = 1$ ,

$$\int_a^b |f(x)| dx \leq |b-a|^{1/q} \left( \int_a^b |f(x)|^p dx \right)^{1/p} \leq |b-a|^{1/q} \|f\|_p.$$

Replace  $f$  with  $f_n - f$ . It is then clear from the above bound that  $L_p$ -convergence  $f_n \rightarrow f$  implies, for any  $p > 1$ ,  $L_1$ -convergence. (The terminology  $L_p$ -convergence is stronger than  $L_1$ -convergence is sometimes used). With a modified bound, we can derive an even stronger proposition, namely that on a bounded interval,  $L_q$ -convergence is stronger than  $L_p$ -convergence whenever  $q > p$ ; try this by yourselves.

If uniform boundedness of the sequence of functions  $f_n$  is given, then there is a constant  $C$  independent of  $n$ , so that  $\|f_n\|_\infty \leq C$ .

Then we can assert that  $|f_n(x) - f(x)| \leq 2C$ , and it then follows that  $L_1$ -convergence implies  $L_p$ -convergence, since

**7.A.21.** Using the Fourier series of the  $2\pi$ -periodic extension of the function  $g(x) = |x|$ , with  $x \in [-\pi, \pi)$ , sum the series

$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^2}.$$

**Solution.** To determine the value of this series, one can successfully apply several known Fourier series. The Fourier series of the function  $g$  coincides with the Fourier series presented in 7.A.11, i.e.,

$$\frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2}.$$

Since this function  $g$  is continuous on  $[-\pi, \pi)$  and  $|\cdot - \pi| = |\pi|$ , we have

$$|x| = \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2}, \quad x \in [-\pi, \pi].$$

Substituting  $x = 0$ , we get  $0 = \frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2}$ . Therefore

$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{\pi^2}{8}.$$

□

**7.A.22.** Determine the convergence and uniform convergence of the Fourier series for the function  $g(x) = e^{-x}$ ,  $x \in [-1, 1)$ .

**Solution.** Again, it is unnecessary to calculate the corresponding Fourier series, since we wish only to check convergence. Let us consider the 2-periodic function  $s(x)$ , defined as follows:

$$s(x) = g(x) = e^{-x}, \quad x \in (-1, 1),$$

$$s(1) = \frac{g(-1) + \lim_{x \rightarrow 1^-} g(x)}{2} = \frac{e + e^{-1}}{2}.$$

According to 7.1.8, this function is the sum of the Fourier series in question. In other words, the Fourier series converges to  $s(x)$ . Moreover, this convergence is uniform on every closed interval which contains none of the points  $2k + 1$ ,  $k \in \mathbb{Z}$ . This follows from the continuity of the functions  $g$  and  $g'$  on  $(-1, 1)$ . On the other hand, the convergence cannot be uniform on any interval  $(c, d)$  such that  $[c, d]$  contains an odd integer. This is because the uniform limit of continuous functions is always a continuous function, and the periodic extension of  $s$  is not continuous at the odd integers. Thus, the series converges to the function  $g$  on  $(-1, 1)$ , yet this convergence is uniform only on the subintervals  $[c, d]$  which satisfy the restriction  $-1 < c < d < 1$ . □

**7.A.23.** Express the function  $g(x) = \pi^2 - x^2$  on the interval  $[-\pi, \pi]$  in the form of a Fourier series. Using this expression, sum the two series



then

$$\int_a^b |f(x) - f_n(x)|^p dx =$$

$$= \int_a^b |f(x) - f_n(x)|^{p-1} |f(x) - f_n(x)| dx$$

$$\leq (2C)^{p-1} \int_a^b |f(x) - f_n(x)| dx$$

which can be written

$$\|f - f_n\|_p \leq (2C)^{1/q} \|f - f_n\|_1^{1/p}.$$

It follows that the  $L_p$ -norms on the space  $S^0[a, b]$  are equivalent with respect to the convergence of uniformly bounded sequences of functions.

**7.1.16. Implications of the Dirichlet condition.** The most difficult (and most interesting) problem is to prove the first statement of the theorem 7.1.8, which in the literature is often referred to as the *Dirichlet condition*, which seems to have been derived as early as in 1824.

We begin by proving how this property of pointwise convergence implies the statements (2) and (3) of the theorem. Without loss of generality, we assume that we are working on the interval  $[-\pi, \pi]$ , i. e. with period  $T = 2\pi$ .

As the first step, we prepare simple bounds for the coefficients of the Fourier series. One bound is of course

$$|a_n| \leq \frac{1}{\pi} \int_{-\pi}^{\pi} |f(x)| dx,$$

and similarly for all the coefficients  $b_n$ . This is because both  $\cos(x)$  and  $\sin(x)$  are bounded by 1 in absolute value. However, if  $f$  is a continuous function in  $S^1[-\pi, \pi]$ , we can integrate by parts, thus obtaining (we write  $a_n(f)$  for the corresponding coefficient of the function  $f$ , and so on)

$$b_n(f) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx$$

$$= \frac{-1}{n\pi} [f(x) \cos(nx)]_{-\pi}^{\pi} + \frac{1}{n\pi} \int_{-\pi}^{\pi} f'(x) \cos(nx) dx$$

$$= \frac{1}{n} a_n(f').$$

Notice  $f(-\pi) = f(\pi)$  by the continuity and thus the boundary term vanishes. Similarly we compute  $a_n(f) = -b_n(f')$ .

Iterating this procedure, we obtain a bound for functions  $f$  in  $S^{k+1}[-\pi, \pi]$  with continuous derivatives up to order  $k$  inclusive

$$|a_n(f)| \leq \frac{1}{n^{k+1}\pi} \int_{-\pi}^{\pi} |f^{(k+1)}(x)| dx,$$

and similarly for  $b_n(f)$ .

Thus we can see that the “smoother” a function is, the more rapidly the Fourier coefficients approach zero. For sufficiently smooth functions  $f$ , the  $n^k$ -multiples of their Fourier coefficients  $a_n$  and  $b_n$  are bounded by the  $L_1$ -norm of their  $k$ -th derivative  $f^{(k)}$ .

$$(1) \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}, \quad \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

**Solution.** We could take advantage of the function  $g$  being even, and calculate the non-zero coefficients  $a_n$  by integration by parts. However, in 7.1.10 the Fourier series for the function  $f(x) = x^2$  on the interval  $[-1, 1]$  is derived. This proves the identity

$$f(x) = \frac{1}{3} + \frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n \cos(n\pi x)}{n^2}, \quad x \in (-1, 1),$$

valid also for  $x = \pm 1$ . By adding  $\pi^2$  and rescaling, it follows

$$g(x) = \pi^2 - \left( \frac{1}{3} + \frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n \cos \frac{n\pi x}{\pi}}{n^2} \right) \pi^2$$

that

$$= \frac{2}{3}\pi^2 + 4 \sum_{n=1}^{\infty} \frac{(-1)^{n+1} \cos(nx)}{n^2}, \quad x \in [-\pi, \pi].$$

Of course, one can also calculate the Fourier series of the function  $g$  directly.

Substituting  $x = 0$  and  $x = \pi$  then gives

$$\pi^2 = \frac{2}{3}\pi^2 + 4 \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2}, \quad \text{i.e.} \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12},$$

and

$$0 = \frac{2}{3}\pi^2 + 4 \sum_{n=1}^{\infty} \frac{(-1)^{n+1}(-1)^n}{n^2}, \quad \text{i.e.} \quad \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

In other words,

$$\begin{aligned} \pi^2 &= 12 \left( 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \dots \right) \\ &= 6 \left( 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \dots \right). \end{aligned}$$

□

We have seen that the Fourier series of a  $2L$ -periodic odd function involves only sine functions, whereas Fourier series of a  $2L$ -periodic even function involves only cosine functions. However, Fourier series are employed in many areas and often we are interested in functions that are only defined on intervals of the form  $[0, L]$ , for some  $L > 0$ . In particular, to get the cosine series expansion of  $f$  we may extend it to  $[-L, L]$  so that the extended function is an even  $2L$ -periodic function. In this case, by the result in 6.B.35 we will have

$$a_n = \frac{2}{L} \int_0^L f(x) \cos\left(\frac{n\pi x}{L}\right) dx, \quad b_n = 0.$$

This gives the cosine extension

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right).$$

Similarly, to get the sine series expansion of  $f$ , we may extend it to  $[-L, L]$  in such a way that the extended function is an odd  $2L$ -periodic function. Then we will have

$$a_n = 0, \quad b_n = \frac{2}{L} \int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx,$$

Let  $f$  be a continuous function in the space  $\mathcal{S}^1[a, b]$ . By the Dirichlet condition, its Fourier series converge pointwise to  $f$ . Then we can assert that

$$\begin{aligned} |s_N(x) - f(x)| &= \left| \sum_{k=N+1}^{\infty} (a_k \cos(kx) + b_k \sin(kx)) \right| \\ &\leq \sum_{k=N+1}^{\infty} (|a_k| + |b_k|). \end{aligned}$$

The right-hand side can further be estimated by the coefficients  $a'_n$  and  $b'_n$  of the derivative  $f'$ . By applying in succession the inequality above, then Hölder's inequality for infinite series (with  $p = q = 2$ ), and then with the arithmetic-geometric inequality, we obtain

$$\begin{aligned} |s_N(x) - f(x)| &\leq \sum_{k=N+1}^{\infty} \frac{1}{k} (|a'_k| + |b'_k|) \\ &\leq \left( \sum_{k=N+1}^{\infty} \frac{1}{k^2} \right)^{\frac{1}{2}} \left( \sum_{k=N+1}^{\infty} (|a'_k|^2 + 2|a'_k||b'_k| + |b'_k|^2) \right)^{\frac{1}{2}} \\ &\leq \left( \sum_{k=N+1}^{\infty} \frac{1}{k^2} \right)^{\frac{1}{2}} \left( \sum_{k=N+1}^{\infty} (2|a'_k|^2 + 2|b'_k|^2) \right)^{\frac{1}{2}} \\ &\leq \sqrt{2} \left( \int_N^{\infty} \frac{1}{x^2} dx \right)^{\frac{1}{2}} \frac{1}{\sqrt{\pi}} \|f'\|_2 \\ &= \left( \frac{\sqrt{2}}{\sqrt{\pi}} \|f'\|_2 \right) \cdot \frac{1}{\sqrt{N}}. \end{aligned}$$

Thus we have obtained not only a proof of the uniform convergence of our series to the anticipated value, but also a bound for the speed of the convergence:

#### UNIFORM CONVERGENCE UNDER THE DIRICHLET CONDITION

If  $f$  is a continuous function in  $\mathcal{S}^1[a, b]$ , then

$$\sup_{x \in \mathbb{R}} |s_N(x) - f(x)| \leq \left( \frac{\sqrt{2}}{\sqrt{\pi}} \|f'\|_2 \right) \cdot \frac{1}{\sqrt{N}}.$$

This proves the statement 7.1.8.(2), supposing the Dirichlet condition 7.1.8.(1) holds.

**7.1.17.  $L_2$ -convergence.** In the next step of our proof, we derive  $L_2$ -convergence of Fourier series. The proof utilizes the common technique of approximation objects which are not continuous by ones which are. We describe it without further details. Interested readers should be able to fill in the gaps by themselves without any difficulties. First, we formulate the statement we need.

**Lemma.** The subset of continuous functions  $f$  in  $\mathcal{S}^0[a, b]$  on a finite interval  $[a, b]$  is a dense subset in this space with respect to the  $L_2$ -norm.



which gives the sine extension

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right),$$

see also at the end of the paragraph 7.1.10. Let us describe a few examples and for further material we refer to Section D.

**7.A.24.** Describe the cosine Fourier series for the sine function  $f(x) = \sin(x)$ , with  $x \in [0, \pi]$ . ○

**7.A.25.** Describe the sine Fourier series for the function  $f(x) = \cos(x)$ , with  $x \in [0, \pi]$ . ○

We end this section with a tasks on the so called “Parseval’s” identity, and some applications of Fourier series.

**7.A.26.** Using Parseval’s identity for Fourier’s orthogonal system (part (3) of the theorem 7.1.5), verify that



$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \frac{\pi^4}{96}.$$

**Solution.** It is imperative to choose an appropriate Fourier series. For instance, consider the Fourier series

$$\frac{\pi}{2} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)x)}{(2n-1)^2},$$

which we obtained for the function  $g(x) = |x|$ ,  $x \in [-\pi, \pi]$  in 7.A.15(b). Parseval’s identity

$$\frac{a_0^2}{2} + \sum_{n=1}^{\infty} a_n^2 + \sum_{n=1}^{\infty} b_n^2 = \frac{2}{T} \int_{x_0}^{x_0+T} (g(x))^2 dx$$

says, substituting for the  $a$ ’s and  $b$ ’s from our particular series, that

$$\frac{\pi^2}{2} + \frac{16}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \frac{1}{\pi} \int_{-\pi}^{\pi} |x|^2 dx = \frac{2}{\pi} \int_0^{\pi} x^2 dx = \frac{2\pi^2}{3},$$

so,

$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \left(\frac{2\pi^2}{3} - \frac{\pi^2}{2}\right) \frac{\pi^2}{16} = \frac{\pi^4}{96}.$$

There are other ways of obtaining this result, see for example (3) at the page 698. We recommend comparing the solutions of this exercise to the previous one.

We started our discussion of Fourier series with the simplest of periodic functions



$$f(t) = a \sin(\omega t + b)$$

for certain constants  $a, \omega > 0, b \in \mathbb{R}$ . They appear as the general solution to the homogeneous linear differential equation

$$(1) \quad y'' + \omega^2 y = 0$$

which arises in mechanics by Newton’s law of force for a moving particle. Recall the brief introduction to the simplest differential equations in 6.2.21 on page 552. Much more follows in Chapter 8.

**SKETCH OF THE PROOF.** Here “dense” means that for any  $g$  in  $S^0[a, b]$  and any  $\varepsilon > 0$ , there is some continuous  $f$  satisfying  $\|f - g\|_2 < \varepsilon$ . We deal with abstract topological concepts like this in the last part of this chapter.

The idea of the proof can be seen easily via the example of approximation of Heaviside’s function  $h$  on the interval  $[-\pi, \pi]$ . We recall that  $h(x) = -1$  for  $x < 0$ , and  $h(x) = 1$  for  $x > 0$ . For every  $\delta$  satisfying  $\pi > \delta > 0$ , we define the function  $f_\delta$  as  $x/\delta$  for  $|x| \leq \delta$  and  $f_\delta(x) = h(x)$  otherwise. All the functions  $f_\delta$  are continuous, in fact, piecewise linear. It can be calculated easily that  $\|h - f_\delta\|_2 \rightarrow 0$  so that  $h$  can be approximated in  $L_2$  norm by a sequence of continuous functions.

All discontinuity points of a general function  $f$  can be catered for in exactly the same way. There are only finitely many of them, and so all of the considered functions are limit points of sequences of continuous functions. □

Now, our proof of the  $L_2$  convergence (under the assumption of the Dirichlet condition) is already simple because for the given function  $f$ , the distance between the partial sums of its Fourier series can be bounded by using a sequence of continuous functions  $f_\varepsilon$  in this way (all norms in this paragraph are the  $L_2$  norms):

$$\|f - s_N(f)\| \leq \|f - f_\varepsilon\| + \|f_\varepsilon - s_N(f_\varepsilon)\| + \|s_N(f_\varepsilon) - s_N(f)\|$$

and the particular summands on the right-hand side can be controlled.

The first one of them is at most  $\varepsilon$ , and according to the uniform convergence for continuous functions (as just proved above), the second summand can be bounded also by  $\varepsilon$ , if  $N$  is big enough. Notice that the third term has the value of the partial sum of the Fourier series for  $f - f_\varepsilon$ . Thus (cf. Theorem 7.1.5(1)),

$$\|f - f_\varepsilon - s_N(f - f_\varepsilon)\| \leq \|f - f_\varepsilon\|.$$

Therefore by the triangle inequality,

$$\|s_N(f - f_\varepsilon)\| \leq \|s_N(f - f_\varepsilon) - f + f_\varepsilon\| + \|f - f_\varepsilon\| \leq 2\|f - f_\varepsilon\|.$$

Altogether,  $\|f - s_N(f)\| \leq 4\varepsilon$ .

This verifies the  $L_2$  convergence of the functions  $s_N(f)$  to  $f$  under the Dirichlet condition, which is we wanted to prove.

**7.1.18. Dirichlet kernel.** Finally, we arrive at the proof of the first statement of theorem 7.1.8. It follows from the definition of the Fourier series  $F(t)$  for a function  $f(t)$ , using its expression with the complex exponential in 7.1.7, that the partial sums  $s_N(t)$  can be written as



$$s_N(t) = \frac{1}{T} \sum_{k=-N}^N \int_{-T/2}^{T/2} f(x) e^{-i\omega kx} e^{i\omega kt} dx,$$



We mention that the function  $f$  has period  $T = 2\pi/\omega$ . In mechanics, one often talks about frequency  $1/T$ . The positive value  $a$  expresses the maximum displacement of the oscillating point from the equilibrium position and it is called the amplitude. The value  $b$  determines the position of the point at the initial time  $t = 0$  and it is called the initial phase, while  $\omega$  is the angular frequency of the oscillation.

Similarly, the function  $z \equiv g(t)$  describes the dependence of voltage upon time  $t$  in an electrical circuit with inductance  $L$  and capacity  $C$  and which is the solution of the differential equation

$$(2) \quad z'' + \omega^2 z = 0.$$

The only difference between the equations (1) and (2) (besides the dissimilar physical interpretation) is the constant  $\omega$ . In the equation (1), there is  $\omega^2 = k/m$  where  $k$  is the proportionality constant and  $m$  is the mass of the point, while in the equation (2), there is  $\omega^2 = (LC)^{-1}$ .

We illustrate how Fourier series can be applied in the theory of differential equations. Consider only the non-homogeneous (compare to (1)) differential equation

$$(3) \quad y'' + a^2 y = f(x)$$

with  $y$  an unknown in variable  $x \in \mathbb{R}$ , with a periodic, continuously differentiable function  $f : \mathbb{R} \rightarrow \mathbb{R}$  on the right-hand side and a constant  $a > 0$ . Let  $T > 0$  be the prime period of the function  $f$  and let its Fourier series on  $[-T/2, T/2]$  be known, i.e. we assume

$$(4) \quad f(x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} \left( A_n \cos \frac{2\pi n x}{T} + B_n \sin \frac{2\pi n x}{T} \right).$$

**7.A.27.** Prove that if the equation (3) has a periodic solution on  $\mathbb{R}$ , then the period of this solution is also a period of the function  $f$ . Further, prove that the equation (3) has a unique periodic solution with period  $T$  if and only if



$$(1) \quad a \neq \frac{2\pi n}{T} \quad \text{for every } n \in \mathbb{N}.$$

**Solution.** Let a function  $y = g(x)$ ,  $x \in \mathbb{R}$ , be a solution of the equation (3) with  $f(x) \not\equiv 0$  and with period  $p > 0$ . In order to substitute the function  $g$  into a second-order differential equation, its second derivative  $g''$  must exist. Since the functions  $g, g', g'', \dots$  share the same period, the function

$$g''(x) + a^2 g(x) = f(x)$$

is also periodic with period  $p$ . In other words, the function  $f$  is periodic as a linear combination of functions with period  $p$ . Thus, we have proved the first statement claiming that  $p = lT$  for a certain  $l \in \mathbb{N}$ .

Suppose that the function  $y = g(x)$ ,  $x \in \mathbb{R}$ , is a periodic solution of the equation (3) with period  $T$  and that it is expressed by a Fourier series as follows:

$$(2) \quad g(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(\omega n x) + b_n \sin(\omega n x)), \quad x \in \mathbb{R},$$

where  $T$  is the period we are working with and  $\omega = 2\pi/T$ . This expression can be rewritten as

$$(1) \quad s_N(t) = \int_{-T/2}^{T/2} K_N(t-x)f(x) dx,$$

where the function

$$K_N(y) = \frac{1}{T} \sum_{k=-N}^N e^{i\omega k y}$$

is called the *Dirichlet kernel*. The sum is a (finite) geometric series with common ratio  $e^{i\omega y}$ . By multiplying by  $e^{i\omega y}$  and then subtracting, we obtain

$$e^{i\omega y} K_N(y) = \frac{1}{T} \sum_{k=-N}^N e^{i\omega(k+1)y}$$

$$(1 - e^{i\omega y})K_N(y) = \frac{1}{T} \left( e^{-iN\omega y} - e^{i(N+1)\omega y} \right).$$

Provided  $\omega y$  is not a multiple of  $2\pi$ , we continue to obtain

$$\begin{aligned} K_N(y) &= \frac{1}{T} \frac{e^{-iN\omega y} - e^{i(N+1)\omega y}}{1 - e^{i\omega y}} \\ &= \frac{1}{T} \frac{1 - e^{-i(N+1/2)\omega y} + e^{i(N+1/2)\omega y}}{e^{i\omega y/2} - e^{-i\omega y/2}} \\ &= \frac{1}{T} \frac{\sin((N+1/2)\omega y)}{\sin(\omega y/2)} \end{aligned}$$

in which the key step was to multiply both the numerator and the denominator by  $-e^{-i\omega y/2}$ . When  $y = 0$ ,  $K_N(0) = \frac{1}{T}(2N+1)$ .

The last expression shows that  $K_N(y)$  is an even function. By l'Hospital's rule, applied at  $y = 0$ , it is continuous everywhere. Since all the partial sums of the series for the constant function  $f(x) = 1$  also equal 1, we obtain from the definition of the Dirichlet kernel, cf. (1), that

$$\int_{-T/2}^{T/2} K_N(x) dx = 1.$$

In the case of periodic functions, the integrals over intervals whose length equals the period are independent of the choice of the end points. Hence, changing the coordinates, we can also use the expression

$$s_N(x) = \int_{-T/2}^{T/2} K_N(y)f(x+y) dy$$

for the partial sums.

Finally, we are fully prepared. First, we consider the case when the function  $f$  is continuous (and piecewise differentiable) at the point  $x$ . We want to prove that in this case, the Fourier series  $F(x)$  for the function  $f$  converges to the value  $f(x)$  at the point  $x$ . We have

$$s_N(x) - f(x) = \int_{-T/2}^{T/2} (f(x+y) - f(x))K_N(y) dy.$$

where  $\omega = 2\pi/T$ . If  $g$  satisfies the equation (3), it has a continuous second derivative on  $\mathbb{R}$ . Therefore, for  $x \in \mathbb{R}$ ,

$$(3) \quad \begin{aligned} g'(x) &= \sum_{n=1}^{\infty} (\omega n b_n \cos(\omega n x) - \omega n a_n \sin(\omega n x)), \\ g''(x) &= \sum_{n=1}^{\infty} (-\omega^2 n^2 a_n \cos(\omega n x) - \omega^2 n^2 b_n \sin(\omega n x)), \end{aligned}$$

Substituting (4), (2) and (3) into (3) yields

$$\begin{aligned} a^2 \frac{1}{2} a_0 + \sum_{n=1}^{\infty} ((-\omega^2 n^2 a_n + a^2 a_n) \cos(n\omega x) + (-\omega^2 n^2 b_n + a^2 b_n) \sin(n\omega x)) \\ = \frac{A_0}{2} + \sum_{n=1}^{\infty} (A_n \cos(n\omega x) + B_n \sin(n\omega x)). \end{aligned}$$

It follows that

$$(4) \quad a^2 \frac{a_0}{2} = \frac{A_0}{2}, \quad \text{that is} \quad a^2 a_0 = A_0,$$

and for  $n \in \mathbb{N}$ ,

$$(5) \quad (-\omega^2 n^2 + a^2) a_n = A_n, \quad (-\omega^2 n^2 + a^2) b_n = B_n.$$

There is exactly one pair of sequences  $\{a_n\}_{n \in \mathbb{N} \cup \{0\}}$ ,  $\{b_n\}_{n \in \mathbb{N}}$  satisfying these conditions if and only if

$$-\omega^2 n^2 + a^2 = -\left(\frac{2\pi n}{T}\right)^2 + a^2 \neq 0 \quad \text{for every } n \in \mathbb{N},$$

i.e., if (1) holds. In this case, the only solution of (3) with period  $T$  is determined by the only solution

$$(6) \quad a_n = \frac{A_n}{-\omega^2 n^2 + a^2}, \quad b_n = \frac{B_n}{-\omega^2 n^2 + a^2}, \quad n \in \mathbb{N}$$

of the system (5). We emphasize that we utilized the uniform convergence of the series in (3).  $\square$

**7.A.28.** Using the solution of the previous problem, find all  $2\pi$ -periodic solutions of the differential equation

$$y'' + 2y = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^2}, \quad x \in \mathbb{R}.$$

**Solution.** The equation is in the form of (3) for  $a = \sqrt{2}$  and the continuously differentiable function

$$f(x) = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^2}, \quad x \in \mathbb{R}$$

with prime period  $T = 2\pi$ . According to problem 7.A.27, the condition  $\sqrt{2} \notin \mathbb{N}$  implies that there is exactly one  $2\pi$ -periodic solution. If we look for it as the value of the series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx)), \quad x \in \mathbb{R},$$

we know also that (see (4) and (6))

$$a_0 = a_n = 0, \quad b_n = \frac{1}{n^2(2-n^2)}, \quad n \in \mathbb{N}.$$

Thus, the given equation has the unique  $2\pi$ -periodic solution

$$y = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^2(2-n^2)}, \quad x \in \mathbb{R}. \quad \square$$

The integrand can be rewritten

$$\begin{aligned} \frac{f(x+y) - f(x)}{\sin(\omega y/2)} \sin((N+1/2)\omega y) &= \\ = \varphi_x(y) (\cos(\omega y/2) \sin(N\omega y) + \sin(\omega y/2) \cos(N\omega y)), \end{aligned}$$

where

$$\varphi_x(y) = \frac{f(x+y) - f(x)}{\sin(\omega y/2)}$$

for  $y \neq 0$ . Moreover, we can compute the right and left limits of  $\varphi_x$  at the point  $y = 0$  using the L'Hospital's rule. Indeed, writing  $f'_{\pm}(x)$  for the right and left derivatives of  $f$ , we arrive at

$$\begin{aligned} \lim_{y \rightarrow 0^+} \varphi_x(y) &= \lim_{y \rightarrow 0^+} \frac{f(x+y) - f(x)}{y} \frac{y}{\sin(\omega y/2)} \\ &= \frac{2}{\omega} f'_+(x), \end{aligned}$$

and similiary for the left limit. Thus,  $\varphi_x(y)$  is piecewise continuous on the entire interval  $[-T/2, T/2]$ .

Next, we rewrite the integral expression for  $s_N - f$  in the form reminding coefficients of Fourier series:

$$\frac{2}{T} \int_{-T/2}^{T/2} (\psi_1(y) \sin(N\omega y) + \psi_2(y) \cos(N\omega y)) dy$$

with the piecewise continuous functions

$$\psi_1(y) = \frac{T}{2} \varphi_x(y) \cos(\omega y/2), \quad \psi_2(y) = \frac{T}{2} \varphi_x(y) \sin(\omega y/2).$$

Indeed, we deal with the Fourier coefficients of the functions  $\psi_1$  and  $\psi_2$ , but they have to converge to zero with  $N \rightarrow \infty$  by virtue of the general theorem on abstract Fourier series, see the Bessel inequality in 7.1.5(2). Hence

$$\lim_{N \rightarrow \infty} \int_{-T/2}^{T/2} \psi_1(y) \sin(N\omega y) dy = 0,$$

and

$$\lim_{N \rightarrow \infty} \int_{-T/2}^{T/2} \psi_2(y) \cos(N\omega y) dy = 0.$$

But this means  $\lim_{N \rightarrow \infty} s_N(x) - f(x) = 0$ , as desired.

Now suppose the function  $f$  has a discontinuity at  $x$ . Without loss of generality, we may assume  $x = 0$  (using constant shift of coordinates otherwise).

Since the function belongs to  $\mathcal{S}^1$ , it is already continuous and differentiable on a neighbourhood of the point  $x = 0$  (outside the point itself). Split  $f$  into its even part  $f_1$  and its odd part  $f_2$ . That is, write  $f(x) = f_1(x) + f_2(x)$ , where

$$f_1(x) = \frac{1}{2}(f(x) + f(-x)) \quad \text{for } x \neq 0$$

$$f_1(0) = \frac{1}{2} \left( \lim_{x \rightarrow 0^+} f(x) + \lim_{x \rightarrow 0^-} f(x) \right)$$

$$f_2(x) = \frac{1}{2}(f(x) - f(-x)).$$

Then the even part  $f_1(x)$  is continuous and piecewise differentiable at the point  $x = 0$  because of the existence of the one-sided limits, and so on a neighbourhood of the point  $x = 0$ .

**B. Integral operators and Fourier Transform**

Next we will focus on problems related with the notion of “convolution”, which is a typical integral operation that applies on two functions  $f, g$  and produces a new one, denote by  $(f \star g)$  and defined by



$$(f \star g)(y) = \int_{-\infty}^{\infty} f(x)g(y - x) dx.$$

The convolution integral is a fundamental concept in various fields, e.g., in statistics, signal processing, computer vision, and other. For its basic features and other basic facts, see 7.2.2.

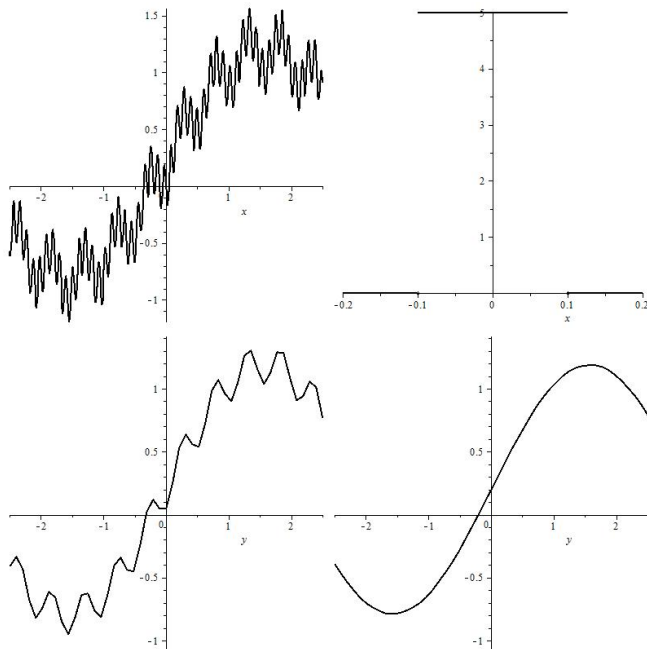
**7.B.1.** Find the convolutions  $f \star g$  and  $f \star h$  of the following functions and in each case check the “smoothing” of  $f$ .

$$f(x) = \sin(x) + \frac{2}{5} [\sin(6x)]^2 - \frac{1}{5} \sin(60x),$$

$$g(x) = \begin{cases} \frac{1}{2\varepsilon} & \text{if } -\varepsilon < x < \varepsilon \\ 0 & \text{otherwise} \end{cases}, \quad \text{where } \varepsilon > 0,$$

$$h(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{where } \mu, \sigma \in \mathbb{R}, \sigma > 0.$$

**Solution.** The function  $g$  is chosen to provide the mean of  $f$  over (small) intervals of length  $2\varepsilon$  and it is normalized so that the integral of  $g$  over all of  $\mathbb{R}$  is one. We should expect some smoothing of the oscillations of the function  $f(x)$ . Drawing the resulting functions by Maple shows:



The graph depicted in the upper left is the original  $f$ , while the graph in the upper right is of  $g$  with  $\varepsilon = 1/10$ . The two

Also  $f_2(0) = 0$ , and the Fourier series for  $f_2$  contains only the terms with  $\sin(n\omega x)$  and thus, vanishes at  $x = 0$ .

Thus we can refer to the previous continuous case and obtain, for the Fourier series  $F(x)$  of the function  $f$ , the equation

$$F(0) = \frac{1}{2} \left( \lim_{x \rightarrow 0^+} f(x) + \lim_{x \rightarrow 0^-} f(x) \right) + 0,$$

which is what we wanted to prove.

If we face more than one jump of  $f$  within the basic period of length  $T$ , we may express  $f$  as a sum of functions with exactly one discontinuity. For example, if points of discontinuity were  $x_0 < x_1$ , both in  $(-T/2, T/2)$ , and writing  $f(y_{\pm}) = \lim_{x \rightarrow y_{\pm}} f(x)$  for the one-sided limits, we can consider  $f = g_1 + g_2$  with

$$g_1(x) = \begin{cases} f(x) & x \in (-T/2, x_0) \\ f(x) - f(x_{0+}) + f(x_{0-}) & x \in (x_0, T/2) \end{cases}$$

$$g_2(x) = \begin{cases} 0 & x \in (-T/2, x_0) \\ f(x_{0+}) - f(x_{0-}) & x \in (x_0, T/2). \end{cases}$$

Clearly  $g_1$  has got the only discontinuity in  $x_1$  while the piecewise constant  $g_2$  jumps at  $x_0$  only. By the already proved behavior, the Fourier series of  $g_1$  and  $g_2$  converges at  $x_0$  to  $f(x_{0-})$  and  $\frac{1}{2}(f(x_{0+}) - f(x_{0-}))$ , respectively. Their sum provides the desired result. Similarly for the other discontinuity point.

This completes the proof. This also completes the proof of the statements (2) and (3) of theorem 7.1.8 where we required that the Dirichlet condition be true.

**2. Integral operators and Fourier transform**

This section provides a few glimpses towards a fascinating and useful area of mathematics exploiting the integration process in many practical ways. Unlike most of this book, we do not provide any general theory and our constructions will be rather illustrated by examples than by rigorous theorems with full proofs.

**7.2.1. Functionals.** In the case of finite-dimensional vector spaces, we can regard the vectors as mappings from a finite set of fixed generators into the space of coordinates. The sums of vectors and the scalar multiples of vectors were then given by the corresponding operations with such functions. We also worked with the vector spaces of functions of a real variable in the same way when their values were scalars (or vectors as well).

The simplest linear mapping  $\alpha$  between vector spaces maps vectors to scalars. It is called a linear form, or a linear functional (in particular when dealing with vector spaces of functions). In finite dimension, it is defined as the sum of products of coordinates  $x_i$  of vectors with fixed values  $\alpha_i = \alpha(e_i)$  at the generators  $e_i$ , i.e. by one-row matrices:

$$(x_1, \dots, x_n)^T \mapsto (\alpha_1, \dots, \alpha_n) \cdot (x_1, \dots, x_n)^T.$$

labels in drawings - proc ne, ale pripsat rukou, popis je ale v textu stejne ?

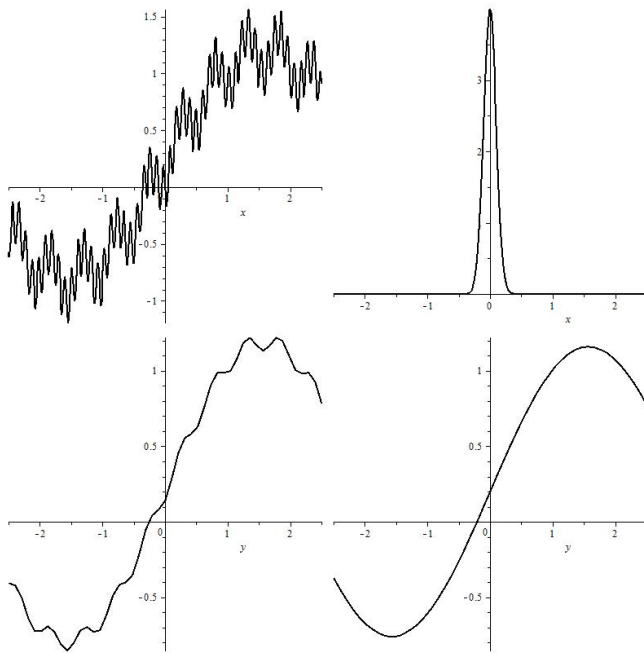
lower graphs show their convolution with (respectively) parameters for  $g$  selected as  $\varepsilon = 1/10$  and  $\varepsilon = 13/50$ . It is straightforward to compute the convolution explicitly, too:

$$f * g(y) = \frac{1}{2\varepsilon} \int_{y-\varepsilon}^{y+\varepsilon} \sin(x) + \frac{2}{5} \sin(6x)^2 - \frac{1}{5} \sin(60x) dx$$

$$= \frac{1}{2\varepsilon} \left[ -\cos(x) - \frac{1}{30} \sin(6x) \cos(6x) + \frac{1}{5}x + \frac{1}{300} \cos(60x) \right]_{y-\varepsilon}^{y+\varepsilon}$$

Similarly, the other function  $h$  is a typical smoothing function which gives much more weight to the values of  $f$  near the point  $y$ , and much less weight to the values of  $f$  further from  $y$ . This is the famous *Gaussian* function. We meet it frequently. Using Maple again, we see that the integral of  $h$  over  $\mathbb{R}$  is one (we prove this in 13.2.8). It is not easy to find the convolution analytically, but Maple can do it approximately. The resulting diagrams are as follows.

As before, the graph depicted in the upper left is the original  $f$ , while the graph in the upper right is of  $h$  with  $\mu = 0$ , and  $\sigma = 1/10$ . Below that, their convolutions are shown with parameters  $\mu = 0$ ,  $\sigma^2 = 1/60$  and (lower right)  $\mu = 0$ ,  $\sigma^2 = 5/60$ .



**7.B.2.** Determine the convolution  $f_1 * f_2$  where

$$f_1(x) = \frac{1}{x} \quad \text{for } x \neq 0$$

$$f_2(x) = \begin{cases} x & \text{for } x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

**Solution.**

More complicated mappings, with values lying in the same space, are given similarly by square matrices. We approach linear operations on spaces of functions in an analogous way.

For simplicity, we work with the real vector space  $\mathcal{S}$  of all piecewise continuous real-valued functions having compact support and defined on the whole  $\mathbb{R}$  or on an interval  $I \neq \varepsilon[a, b]$ . Linear mappings  $\mathcal{S} \rightarrow \mathbb{R}$  are called (real) *linear functionals*. Such functionals can be defined in many diverse ways. For example, by evaluating the function's values (or its derivatives') at some fixed points or in terms of integration.

We can, for instance consider the functional  $L$  given by evaluating the function at a sole fixed point  $x_0 \in I$ , i. e.,

$$L(f) = f(x_0).$$

Or, we can have the functional given by integration of the product with a fixed function  $g(x)$ , i.e.,

$$L(f) = \int_a^b f(x)g(x) dx.$$

The function  $g(x)$  in the previous example is a function which weighs the particular values representing the function  $f(x)$  in the definition of the Riemann integral and the functional  $L$  is a perfect analogue of the linear forms on finite dimensional vector spaces mentioned above. The simplest case of such a functional is, of course, the Riemann integral itself, i. e. the case of  $g(x) = 1$  for all points  $x$ .

A good example is given by

$$g(x) = \begin{cases} 0 & \text{if } |x| \geq \varepsilon \\ \frac{1}{2\varepsilon} & \text{if } |x| < \varepsilon. \end{cases}$$

for any  $\varepsilon > 0$ . The integral of the function  $g$  over  $\mathbb{R}$  equals one, and our linear functional can be perceived as a (uniform) averaging of the values of the function  $f$  over the  $\varepsilon$ -neighbourhood of the origin. Similarly, we can work with the function

$$g(x) = \begin{cases} 0 & \text{if } |x| \geq \varepsilon \\ e^{\frac{1}{x^2 - \varepsilon^2} + \frac{1}{\varepsilon^2}} & \text{if } |x| < \varepsilon \end{cases}$$

which we used in the paragraph 6.1.9. This function is smooth on  $\mathbb{R}$  with compact support on the interval  $(-\varepsilon, \varepsilon)$ .

Our functional has the meaning of a weighted combination of the values, but this time, the weights of the input values decrease rapidly as their distance from the origin increases. The integral of  $g$  over  $\mathbb{R}$  is finite, but it is not equal to one. Dividing  $g$  by this integral would lead to a functional which would have the meaning of a non-uniform averaging of a given function  $f$ .

Another example is the Gaussian function

$$g(x) = \frac{1}{\sqrt{\pi}} e^{-x^2},$$

which also has its integral over  $\mathbb{R}$  equal to one (we verify this later). This time, all the input values  $x$  in the corresponding "average" have a non-zero weight, yet this weight becomes insignificantly small as the distance from the origin increases.

$$(f_1 * f_2)(t) = \int_{-\infty}^{\infty} f_1(x) f_2(t-x) dx = \int_{-\infty}^{\infty} \frac{1}{x} f_2(t-x) dx.$$

Since  $f_2(x) = 0$  outside the interval  $[-1, 1]$ , necessarily  $-1 \leq t-x \leq 1$  that is,  $t-1 \leq x \leq t+1$ . So

$$(f_1 * f_2)(t) = \int_{t-1}^{t+1} \frac{1}{x} (t-x) dx = t \int_{t-1}^{t+1} \frac{1}{x} dx - 2.$$

This last integral is improper if  $t-1 \leq 0 \leq t+1$ . For  $t$  outside that interval, the integration gives

$$(f_1 * f_2)(t) = t \ln \left| \frac{t+1}{t-1} \right| - 2.$$

If instead,  $-1 < t < 1$ , we can for small  $\varepsilon > 0$ , replace  $\int_{t-1}^{t+1} \frac{1}{x} dx$  with

$$\int_{\varepsilon}^{t+1} \frac{1}{x} dx + \int_{t-1}^{-\varepsilon} \frac{1}{x} dx$$

which computes to  $\ln |t+1| - \ln |\varepsilon| + \ln |-\varepsilon| - \ln |t-1|$ . The terms in  $\varepsilon$  cancel, so when we take the limit  $\varepsilon \rightarrow 0$ , we obtain the same answer for the integral as before.

Thus

$$(f_1 * f_2)(t) = t \ln \left| \frac{t+1}{t-1} \right| - 2$$

for all values of  $t$  except for  $t = 1$ , or for  $t = -1$ . □

We calculate the convolution of two functions both of which have a bounded support.

**7.B.3.** Determine the convolution  $f_1 * f_2$  where

$$f_1(x) = \begin{cases} 1-x^2 & \text{for } x \in [-1, 1], \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(x) = \begin{cases} x & \text{for } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

**Solution.** Since the integrand is zero when  $f_1(x) = 0$ ,

$$f_1 * f_2(t) = \int_{-\infty}^{\infty} f_1(x) f_2(t-x) dx$$

$$= \int_{-1}^1 (1-x^2) f_2(t-x) dx$$

But the integrand is also zero when  $f_2(t-x) = 0$ , so we need  $0 \leq t-x \leq 1$  ie.  $t-1 < x < t$  for the integrand to be non zero. So for a non zero value of  $f_1 * f_2(t)$ , we integrate over the intersection of the intervals  $[t-1, t]$  and

**7.2.2. Function convolution.** Integral functionals from the previous paragraph can easily be modified to obtain a “streamed averaging” of the values of a given function  $f$  near a given point  $y \in \mathbb{R}$ :



$$L_y(f) = \int_{-\infty}^{\infty} f(x) g(y-x) dx$$

CONVOLUTION OF FUNCTIONS OF A REAL VARIABLE

The free parameter  $y$  in the definition of the functional  $L_y(f)$  can be perceived as a new independent variable, and our operation  $L_y$  actually maps functions to functions again,  $f \mapsto \tilde{f}$ :

$$\tilde{f}(y) = L_y(f) = \int_{-\infty}^{\infty} f(x) g(y-x) dx.$$

This operation is called the *convolution of functions*  $f$  and  $g$ , denoted  $f * g$ .

The convolution is usually defined for real or complex valued functions on  $\mathbb{R}$  with compact support, or those with decay fast enough to ensure the integrability.

By the transformation  $t = z - x$ , we can easily calculate that

$$(f * g)(z) = \int_{-\infty}^{\infty} f(x) g(z-x) dx$$

$$= - \int_{\infty}^{-\infty} f(z-t) g(t) dt = (g * f)(z).$$

Thus the convolution, considered as a binary operation

$$* : \mathcal{S}_c \times \mathcal{S}_c \rightarrow \mathcal{S}_c$$

of pairs of functions having compact support, is commutative.

Similarly, convolutions can be considered with integration over a finite interval; we only have to guarantee that the functions participating in them are well-defined. In particular, this can be done for periodic functions, integrating over an interval whose length equals the period.

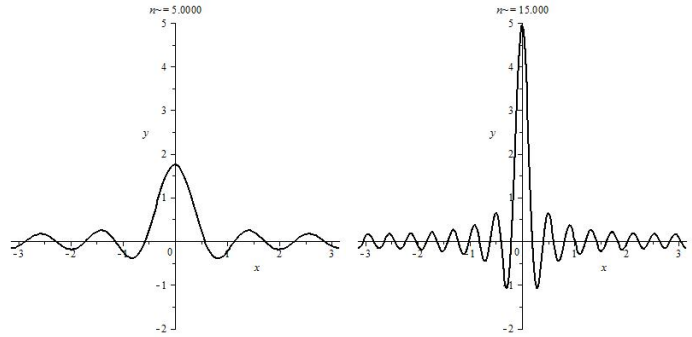
Convolution is an extraordinarily useful tool for modeling the way in which we observe the data of an experiment or the influence of a medium through which information is transferred. For instance, an analog audio or video signal affected by noise. The input value  $f$  is the transferred information. The function  $g$  is chosen so that it would express the influence of the medium or the technical procedure used for the signal processing or the processing of any other data.

**7.2.3. Gibbs phenomenon.** Actually, we have already seen a very useful case of convolution. In paragraph 7.1.18, we interpreted the partial sum of the Fourier series for a function  $f$  as a convolution with the Dirichlet kernel  $K_N(y) = \sum_{-T/2}^{T/2} e^{i\omega_k y}$ . The figure shows this convolution kernel with  $N = 5$  and  $N = 15$ .



$[-1, 1]$ . Consequently,

$$\begin{aligned} (f_1 * f_2)(t) &= 0, \text{ if } t > 2 \\ &= \int_{t-1}^1 (1-x^2)(t-x) dx = 4t/3 - t^2 + t^4/12, \\ &\hspace{10em} 1 \leq t \leq 2 \\ &= \int_{t-1}^t (1-x^2)(t-x) dx = -t^2/2 + 1/4 + 2t/3, \\ &\hspace{10em} 0 \leq t \leq 1 \\ &= \int_{-1}^t (1-x^2)(t-x) dx \\ &= -t^4/12 + t^2/2 + 1/4 + 2t/3, \\ &\hspace{10em} -1 \leq t \leq 0 \\ &= 0, \text{ if } t < -1. \end{aligned}$$



Notice that instead of integrals over the entire real line, we employ the integration over the basic period  $T$  of the periodic functions in question.

This interpretation allows us to explain the Gibbs phenomenon mentioned in paragraph 7.1.9. The point is that we know well the behaviour of the Dirichlet kernel near to the origin and thus, taken into account that the function  $f$  is bounded over the whole period and has all one-side limits of values and derivatives at each point of discontinuity, the effect of the convolution must be quite local. Consequently this leads to verification that the convolution with the Dirichlet kernel in the point  $x$  of jump of  $f$  behaves the same way as we computed explicitly for the Heaviside function at  $x = 0$ . There the overshooting by the Fourier sums can be computed explicitly and this explains the Gibbs effect in general.

We do not provide more details here. Readers may either work them out themselves (as a truly nontrivial exercise) or look them up in the literature.

**7.B.4.** Determine the convolution  $f_1 * f_2$  of the functions

$$\begin{aligned} f_1 &= \begin{cases} 1-x & \text{for } x \in [-2, 1], \\ 0 & \text{otherwise,} \end{cases} \\ f_2 &= \begin{cases} 1 & \text{for } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The next topic is the Fourier transform, which is another example of an integral operator. This time the kernel  $e^{-i\omega t}$  is complex (see 7.2.5 for the terminology). Thus the values on real functions are complex functions in general, see 7.2.5. This is a basic operation in mathematics, allowing the time and frequency analysis of signals and also the transitions between local and global behaviour.

**7.B.5.** Fix  $\Omega > 0$ . Recall that  $\text{sgn } t = 1$  if  $t > 0$ , and  $\text{sgn } t = -1$  if  $t < 0$ ,  $\text{sgn } 0 = 0$ .

Find the Fourier transform  $\mathcal{F}(f)$  and the inverse Fourier transform  $\mathcal{F}^{-1}$  of the functions:

- (a)  $f(t) = \text{sgn } t$  if  $t \in (-\Omega, \Omega)$ , and zero otherwise.
- (b)  $f(t) = 1$  if  $t \in (-\Omega, \Omega)$  and zero otherwise.

**Solution.** The case (a).

The Fourier transform of the given function is

$$\begin{aligned} \mathcal{F}(f)(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \text{sgn } t (\cos(\omega t) - i \sin(\omega t)) dt \\ &= \frac{1}{\sqrt{2\pi}} \left( \int_0^{\Omega} (\cos(\omega t) - i \sin(\omega t)) dt - \int_{-\Omega}^0 (\cos(\omega t) - i \sin(\omega t)) dt \right). \end{aligned}$$

Since  $\cos$  and  $\sin$  are respectively even and odd functions,

**7.2.4. Integral operators.** In general, integral operators can depend on any number of values and derivatives of the function in its argument. For example, considering a function  $F$  depending on  $k + 2$  free arguments,

$$L(f)(y) = \int F(y, x, f(x), f'(x), \dots, f^{(k)}(x)) dx.$$

Convolution is one of many examples of a special class of such operators on spaces of functions

$$L(f)(y) = \int_a^b f(x)k(y, x) dx.$$

The function  $k(y, x)$ , dependent on two variables,

$$k : \mathbb{R}^2 \rightarrow \mathbb{R},$$

is called the *kernel of the integral operator*  $L$ .

The theory of integral operators is very useful and interesting. We focus only on an extraordinarily important special case, namely the *Fourier transform*  $\mathcal{F}$ , which has deep connections with Fourier series.

**7.2.5. Fourier transform.** Recall that a function  $f(t)$ , given by its converging Fourier series, equals

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{i\omega_n t},$$

$$\mathcal{F}(f)(\omega) = \frac{2}{\sqrt{2\pi}} \int_0^{\Omega} -i \sin(\omega t) dt = \frac{2i}{\sqrt{2\pi}} \left[ \frac{\cos(\omega t)}{\omega} \right]_0^{\Omega} \\ = i \sqrt{\frac{2}{\pi}} \frac{\cos(\omega\Omega) - 1}{\omega}.$$

The inverse Fourier transform is given by almost the same integral, with the kernel  $e^{i\omega x}$  instead of  $e^{-i\omega x}$ . The integration is in the frequency domain with variable  $\omega$ . Thus, the only difference in the result is the sign:

$$\mathcal{F}^{-1}(f)(t) = \frac{2}{\sqrt{2\pi}} \int_0^{\Omega} i \sin(\omega t) d\omega = \frac{-2i}{\sqrt{2\pi}} \left[ \frac{\cos(\omega t)}{t} \right]_0^{\Omega} \\ = i \sqrt{\frac{2}{\pi}} \frac{1 - \cos(t\Omega)}{t}.$$

Case (b) is computed similarly:

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} (\cos(\omega t) - i \sin(\omega t)) dt \\ = \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \cos(\omega t) dt = \frac{1}{\sqrt{2\pi}} \left[ \frac{\sin(\omega t)}{\omega} \right]_{-\Omega}^{\Omega} = \sqrt{\frac{2}{\pi}} \frac{\sin(\omega\Omega)}{\omega}.$$

The latter expression is often expressed by means of the function  $\text{sinc}(t) = \sin(t)/t$

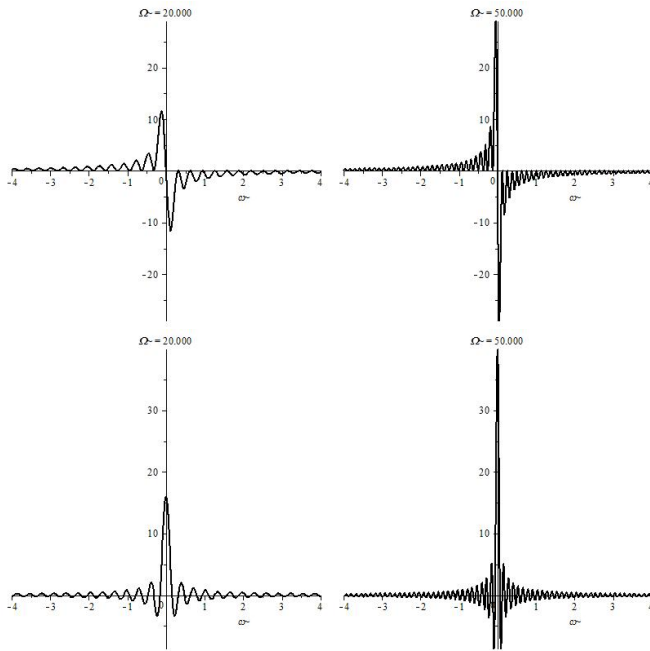
$$\mathcal{F}(f)(\omega) = \frac{2\Omega}{\sqrt{2\pi}} \text{sinc}(\omega\Omega).$$

Here, the inverse Fourier transform has exactly the same result, because the sign change in the kernel does not affect the real part. Thus we only need to interchange the time and frequency variables:

$$\mathcal{F}^{-1}(f)(t) = \frac{2\Omega}{\sqrt{2\pi}} \text{sinc}(t\Omega).$$

□

The results are:



The first two diagrams below show the imaginary values of the Fourier image of the signum function from 7.B.5(a) with  $\Omega = 20$  and  $\Omega = 50$ . The next two diagrams do the same for the characteristic function of the interval  $|x| < \Omega$

where the numbers  $c_n$  are complex Fourier coefficients, and  $\omega_n = n2\pi/T$  with period  $T$ , see paragraph 7.1.7.

After fixing  $T$ , the expression  $\Delta\omega = 2\pi/T$  describes the change of the frequency caused by  $n$  being increased by one. Thus it is just the discrete step by which we change the frequencies when calculating the coefficients of the Fourier series. The coefficient  $1/T$  in the formula

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(x) e^{-i\omega_n x} dx$$

then equals  $\Delta\omega/2\pi$ , so the series for  $f(t)$  can be rewritten as

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{1}{2\pi} \left( \Delta\omega \int_{-T/2}^{T/2} f(x) e^{-i\omega_n x} dx \right) e^{i\omega_n t}.$$

Now imagine the values  $\omega_n$  for all  $n \in \mathbb{Z}$  as the chosen representatives for small intervals  $[\omega_n, \omega_{n+1}]$  of length  $\Delta\omega$ . Then, our expression in the big inner parentheses in the previous formula for  $f(t)$ , restricting the sum to  $n \in [-N, N]$ , describes the summands of the Riemann sums for the integrals



$$\frac{1}{2\pi} \int_{-N}^N g(\omega) e^{i\omega t} d\omega,$$

where  $g(\omega)$  is a function which takes, at the points  $\omega_n$ , the values

$$g(\omega_n) = \int_{-T/2}^{T/2} f(x) e^{-i\omega_n x} dx.$$

Considering the limit for  $N \rightarrow \infty$  leads to the improper integral of  $g(\omega) e^{i\omega t}$ .

If the function  $f(x) e^{-i\omega_n x}$  is integrable over the entire  $\mathbb{R}$  (this is always the case if  $f$  is piecewise continuous with a compact support) then letting  $T \rightarrow \infty$ , the norm  $\Delta\omega$  of our subintervals in the Riemann sum decreases to zero. We obtain the integral

$$g(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx.$$

The previous reasonings indicates that there should be a large set of functions  $f$  on  $\mathbb{R}$  (e.g. all those with  $|f|$  integrable over  $\mathbb{R}$ ) for which we can define a pair of mutually inverse integral operators:



from 7.B.5(b). The longer the interval with the constant values is, the more the image is concentrated around the origin.

*We can always use directly the simpler version of the transform for the odd and even functions. If the argument  $f$  is odd, then only the sine part of the formula contributes and its Fourier transform is*



$$\mathcal{F}(f)(\omega) = \frac{-2i}{\sqrt{2\pi}} \int_0^{\infty} f(t) \sin(\omega t) dt.$$

Similarly, for even functions  $f$

$$\mathcal{F}(f)(\omega) = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} f(t) \cos(\omega t) dt.$$

*In particular, the odd functions have pure imaginary images while the images of the even functions are real. More generally, every real function  $f$  decomposes in to its odd and even parts  $f = f_{\text{even}} + f_{\text{odd}}$  and the real and imaginary components of the Fourier image  $\tilde{f}$  are the images of these two parts, respectively.*

**7.B.6.** Discover how the Fourier transform and its inverse behave under the translation  $\tau_a$  in the variable,  $\tau_a f(x) = f(x+a)$ , and the phase shift  $\varphi_a$  defined as  $\varphi_a f(x) = e^{iax} f(x)$ , always with  $a \in \mathbb{R}$ .

**Solution.** Evaluate the compositions  $\mathcal{F} \circ \tau_a$  and  $\mathcal{F} \circ \varphi_a$ . This is easy:

$$\begin{aligned} \mathcal{F} \circ \tau_a f(\omega) &= \int_{-\infty}^{\infty} f(t+a) e^{-i\omega t} dt \\ &= \int_{-\infty}^{\infty} f(x) e^{-i\omega(x-a)} dx = e^{ia\omega} \mathcal{F}f(\omega). \end{aligned}$$

$$\begin{aligned} \mathcal{F} \circ \varphi_a f(\omega) &= \int_{-\infty}^{\infty} f(t) e^{iat} e^{-i\omega t} dt \\ &= \int_{-\infty}^{\infty} f(t) e^{-i(\omega-a)t} dt = \mathcal{F}f(\omega-a). \end{aligned}$$

Thus is proved the formulae

$$\mathcal{F} \circ \tau_a = \varphi_a \circ \mathcal{F}, \quad \mathcal{F} \circ \varphi_a = \tau_{-a} \circ \mathcal{F}.$$

Similarly,

$$\mathcal{F}^{-1} \circ \tau_a = \varphi_{-a} \circ \mathcal{F}^{-1}, \quad \mathcal{F}^{-1} \circ \varphi_a = \tau_a \circ \mathcal{F}^{-1}.$$

□

*The next problem displays the behaviour of the Fourier transform on the Gaussian function. This is a rare example, where the time and frequency forms are very similar. Again, we see the feature of exchanging the local and global properties in the time and frequency domains.*



**7.B.7.** Compute the Fourier transform  $\mathcal{F}(f)$  of the function

$$f(t) = e^{-at^2}, \quad t \in \mathbb{R},$$

where  $a > 0$  is a fixed parameter.

**Solution.** The task is to calculate

#### FOURIER TRANSFORM

For every real or complex valued function  $f$  on  $\mathbb{R}$  for which the following integrals exist (e.g. all those piecewise continuous with compact support or fast decay), we define

$$\begin{aligned} \mathcal{F}(f)(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \\ \mathcal{F}^{-1}(f)(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\omega) e^{i\omega t} d\omega. \end{aligned}$$

The function  $\tilde{f}(\omega) = \mathcal{F}(f)(\omega)$  is called the *Fourier transform* of the function  $f$ .

The previous ideas indicate that  $\mathcal{F}^{-1}(\mathcal{F}(f))(t) = f(t)$ , whenever both the improper (Riemann) integrals exist.

This says that the Fourier transform  $\mathcal{F}$  has an inverse operation  $\mathcal{F}^{-1}$ , which is called the *inverse Fourier transform*.

Notice, the Fourier transform and its inverse are integral operators with almost identical kernels

$$k(\omega, t) = e^{\pm i\omega t}.$$

The right choice of the function space (and integral type) ensuring the existence of Fourier transform and its inverse is a very subtle question. This is the main reason why we do not formulate theorems with formal proofs and just touch the problems based on examples here.

We hope this will motivate the readers to find more detailed answers in specialized literature.

**7.2.6. Simple properties.** The Fourier transform changes



the local and global behaviour of functions in an interesting way. We begin with a simple example in which there is a function  $f(t)$  which is transformed to the indicator function of the interval  $[-\Omega, \Omega]$ , i. e.,  $\tilde{f}(\omega) = 0$  for  $|\omega| > \Omega$ , and  $\tilde{f}(\omega) = 1$  for  $|\omega| \leq \Omega$ .

The inverse transform  $\mathcal{F}^{-1}$  gives

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} e^{i\omega t} d\omega = \frac{1}{\sqrt{2\pi}} \left[ \frac{1}{it} e^{i\omega t} \right]_{-\Omega}^{\Omega} \\ &= \frac{2}{\sqrt{2\pi} 2i} (e^{i\Omega t} - e^{-i\Omega t}) \\ &= \frac{2\Omega}{\sqrt{2\pi}} \frac{\sin(\Omega t)}{\Omega t}. \end{aligned}$$

Thus, except for a multiplicative constant and the scaling of the input variable, it is the very important function  $f(x) = \text{sinc}(t) = \frac{\sin t}{t}$ .

This function is not that easy to handle. Our construction of the Fourier transform suggests that  $\mathcal{F}(f)$  should be the indicator function of the interval  $[-\Omega, \Omega]$ . We verify this by direct computation in ???. Now, just notice that the sinc function is not integrable in absolute value over  $\mathbb{R}$ . It rather behaves as an alternating series of numbers.

But the local behavior of  $f$  is easy to see. Calculation of the limit at zero, by l'Hospital's rule or otherwise, gives  $f(0) = 2\Omega(2\pi)^{-1/2}$ . The closest zero points are at



$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-at^2} e^{-i\omega t} dt.$$

A standard trick is to transform the problem into one of solving a (simple) differential equation. Hence:

Differentiating (with respect to  $\omega$ ) and then integrating by parts gives

$$\begin{aligned} (\mathcal{F}(f)(\omega))' &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -it e^{-at^2} e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \left( \lim_{t \rightarrow \infty} \frac{i}{2a} e^{-at^2 - i\omega t} - \lim_{t \rightarrow -\infty} \frac{i}{2a} e^{-at^2 - i\omega t} \right. \\ &\quad \left. - \int_{-\infty}^{\infty} \frac{i(-i\omega)}{2a} e^{-at^2} e^{-i\omega t} dt \right) \\ &= \frac{1}{\sqrt{2\pi}} \left( \frac{i}{2a} \lim_{t \rightarrow \infty} e^{-at^2} - \frac{i}{2a} \lim_{t \rightarrow -\infty} e^{-at^2} \right. \\ &\quad \left. - \int_{-\infty}^{\infty} \frac{\omega}{2a} e^{-at^2} e^{-i\omega t} dt \right) \\ &= -\frac{\omega}{2a} \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-at^2} e^{-i\omega t} dt \right) = -\frac{\omega}{2a} \mathcal{F}(f)(\omega). \end{aligned}$$

Therefore  $y(\omega) = \mathcal{F}(f)(\omega)$  satisfies the differential equation

$$\frac{dy}{d\omega} = -\frac{\omega}{2a} y, \quad \text{i.e.} \quad \frac{1}{y} dy = -\frac{\omega}{2a} d\omega,$$

unless  $y$  equals zero ( $y \equiv 0$  is a solution of the equation). Integration yields

$$\ln |y| = -\frac{\omega^2}{4a} - C, \quad \text{i.e.} \quad y = Ke^{-\frac{\omega^2}{4a}},$$

where  $C$  and  $K$  are constants. All solutions (including the zero solution) of the differential equation are given by the function

$$y(\omega) = Ke^{-\frac{\omega^2}{4a}}, \quad K \in \mathbb{R}.$$

To find  $K$ , begin with the well known fact (proved in ??)

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi},$$

to obtain

$$\int_{-\infty}^{\infty} e^{-at^2} dt = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{\sqrt{a}}.$$

Therefore,  $\mathcal{F}(f)(0) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{\sqrt{a}} = \frac{1}{\sqrt{2a}}$  and  $\mathcal{F}(f)(0) = Ke^0 = K$ . So  $K = \frac{1}{\sqrt{2a}}$  and

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2a}} e^{-\frac{\omega^2}{4a}}.$$

□

**7.B.8.** Determine the Fourier transform image of the Gaussian function

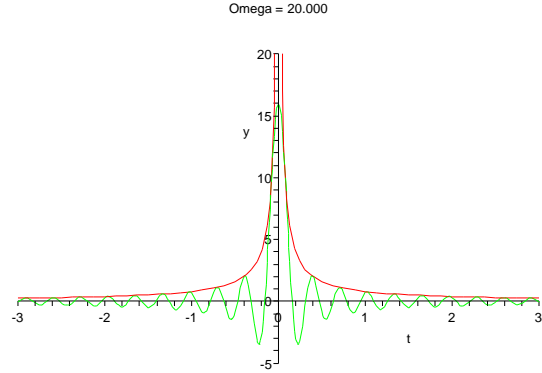
$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

**Solution.** Use the result of the previous problem with  $a = \frac{1}{2\sigma^2}$  and the composition with the variable shift  $\tau_a$  from the last but one problem. It follows that

$$\mathcal{F}(f)(\omega) = \frac{1}{\sqrt{2\pi}} e^{i\mu\omega} e^{-\sigma^2 \frac{\omega^2}{2}}.$$

□

$t = \pm\pi/\Omega$  and the function drops in value to zero quite rapidly away from the origin  $x = 0$ . This function is shown in the diagram by a wavy curve for  $\Omega = 20$ . Simultaneously, the area where our function  $f(t)$  keeps waving more rapidly as  $\Omega$  increases is also depicted by a curve.



The Fourier transform of the indicator function of the interval  $[-\Omega, \Omega]$  is also the above function  $f$ , see the explicit computation in 7.B.5 (the even function sinc does not see the change of sign in the kernel of the transform). Clearly, this function  $f$  takes significant positive values near zero, and the value taken at zero is a fixed multiple of  $\Omega$ . Therefore, as  $\Omega$  increases, the function  $f$  concentrates more and more near the origin.

Next, we derive the Fourier transform of the derivative  $f'(t)$  for a function  $f$ . We continue to suppose that all the integration makes sense (e.g.  $f$  has compact support), so that both  $\mathcal{F}(f')$  and  $\mathcal{F}(f)$  exist. By integration by parts,

$$\begin{aligned} \mathcal{F}(f')(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(t) e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} [e^{-i\omega t} f(t)]_{-\infty}^{\infty} + \frac{i\omega}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \\ &= i\omega \mathcal{F}(f)(\omega). \end{aligned}$$

Thus the Fourier transform converts the (limit) operation of differentiation to the (algebraic) operation of a multiplication by the variable. Of course, this procedure can be iterated, to obtain

$$\mathcal{F}(f^{(n)}) (\omega) = -\omega^2 \mathcal{F}(f), \dots, \mathcal{F}(f^{(n)}) = i^n \omega^n \mathcal{F}(f).$$

**7.2.7. The relation to convolutions.** There is another extremely important property to consider, namely the relation between convolutions and Fourier transforms.

We shall calculate the Fourier transform of the convolution  $h = f * g$ , where we again assume that all the integrals exist (e.g. assuming the functions to be piecewise continuous with compact supports).

Recall that we may change the order of integration, see 6.3.12. Then we change variable by the substitution  $t - x = u$ .

As mentioned, the most typical use of the Fourier transform is to analyse the frequencies in a signal. The next problem reveals the reason. For technical reasons we cut the signal by multiplication with the characteristic function  $h_\Omega$  of the interval  $(-\Omega, \Omega)$ .



**7.B.9.** Find the Fourier transform of the functions

$$f(t) = h_\Omega(t) \cos(nt), \quad g(t) = h_\Omega(t) \sin(nt).$$

**Solution.** By definition

$$\begin{aligned} \mathcal{F}(f)(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \cos(nt) e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\Omega}^{\Omega} \frac{1}{2} (e^{int} + e^{-int}) e^{-i\omega t} dt \\ &= \frac{1}{2\sqrt{2\pi}} \left[ \frac{1}{i(n-\omega)} e^{i(n-\omega)t} \right]_{-\Omega}^{\Omega} \\ &\quad + \frac{1}{2\sqrt{2\pi}} \left[ \frac{-1}{i(n+\omega)} e^{-i(\omega+n)t} \right]_{-\Omega}^{\Omega} \\ &= \frac{\Omega}{\sqrt{2\pi}} (\text{sinc}((n-\omega)\Omega) + \text{sinc}((n+\omega)\Omega)). \end{aligned}$$

The same computation leads to the image of the sine signal, the only difference is one minus sign and an additional  $i$  in the formula:

$$\mathcal{F}(g)(\omega) = i \frac{\Omega}{\sqrt{2\pi}} (-\text{sinc}((n-\omega)\Omega) + \text{sinc}((n+\omega)\Omega))$$

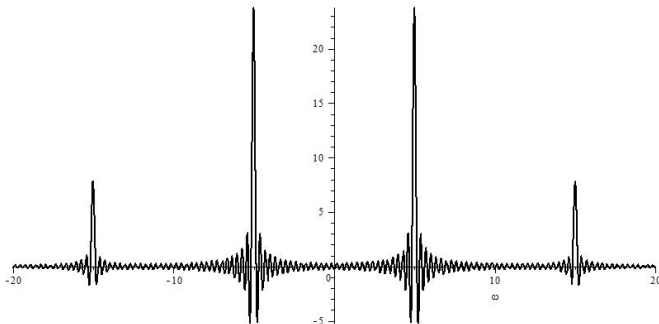
□

**7.B.10.** Find the Fourier transform of the superposition of the  $\cos(nt)$  signals over the interval  $(-\Omega, \Omega)$ ,

$$f(t) = h_\Omega(t)(3 \cos(5t) + \cos(15t)).$$

What happens, if  $\Omega \rightarrow \infty$ ?

**Solution.** The Fourier transform is linear over scalars, thus we simply add the corresponding images from the previous problem with  $n = 1$  and  $n = 3$ , multiplied by the proper coefficients. The illustration of the image with  $\Omega = 20$  is



Each of the peaks behaves like the Fourier image of the characteristic function  $h_\Omega$ , shifted to the frequencies.

The result is

$$\begin{aligned} \mathcal{F}(h)(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} f(x)g(t-x) dx \right) e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \left( \int_{-\infty}^{\infty} g(t-x) e^{-i\omega t} dt \right) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \left( \int_{-\infty}^{\infty} g(u) e^{-i\omega(u+x)} du \right) dx \\ &= \frac{1}{\sqrt{2\pi}} \left( \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx \right) \cdot \left( \int_{-\infty}^{\infty} g(u) e^{-i\omega u} du \right) \\ &= \sqrt{2\pi} \mathcal{F}(f) \cdot \mathcal{F}(g) \end{aligned}$$

A similar calculation shows that the Fourier transform of a product is the convolution of the transforms, up to a multiplicative constant. In fact,

$$\mathcal{F}(f \cdot g) = \frac{1}{\sqrt{2\pi}} \mathcal{F}(f) * \mathcal{F}(g).$$

As we mentioned above, the convolution  $f * g$  often models the process of the observation of some quantity  $f$ . Using the Fourier transform and its inverse, the original values of this quantity are easily recognised if the convolution kernel  $g$  is known. We just calculate  $\mathcal{F}(f * g)$  and divide it by the image  $\mathcal{F}(g)$ . This yields the Fourier transform of the original function  $f$ , which can be obtained explicitly using the inverse Fourier transform. This is sometimes called *deconvolution*.

In real applications, the procedure often cannot be that straightforward since the Fourier image of the known convolution kernel might have zero values and therefore we hardly can divide by it as above. For example, take the convolution kernel  $\text{sinc}(t)$  whose image is an indicator function of some finite interval. So we need some more cunning techniques and there is a vast literature on them.

**7.2.8. The  $L_2$ -norm.** Now we are in position to verify that



the Fourier transform is actually an isometry with respect to the  $L_2$ -norm. The Fourier transform exists for all functions in  $L_1$ , i.e., with integrable absolute values. A simple observation reveals that  $\mathcal{F} : L_2 \rightarrow L_\infty$  satisfies

$$\|\mathcal{F}(f)\|_\infty \leq \|f\|_1.$$

Indeed,  $|e^{-i\omega x}| = 1$  and thus,  $|\mathcal{F}(f)(\omega)| \leq \int_{-\infty}^{\infty} |f(x)| dx$  for all  $\omega$ .

Now, assume  $f, g$  are two functions in  $L_2$  and write  $\hat{g}$  for the function  $\hat{g}(t) = \overline{g(-t)}$ . Notice

$$(f * \hat{g})(t) = \int_{-\infty}^{\infty} f(x) \hat{g}(t-x) dx = \int_{-\infty}^{\infty} f(x) \overline{g(x-t)} dx.$$

In particular, the scalar product is given by the formula

$$\langle f, g \rangle = (f * \hat{g})(0).$$

If  $\Omega$  increases to infinity, the image  $\tilde{f}$  has four peaks at the same positions corresponding to the frequencies  $\pm 5$  and  $\pm 15$ . But they become narrower and sharper. In the limit, this is no longer a function since the width of the peaks becomes zero. This is usually written

$$\mathcal{F}(\cos(nt))(\omega) = \sqrt{\frac{\pi}{2}}(\delta(n - \omega) + \delta(n + \omega))$$

with the special case

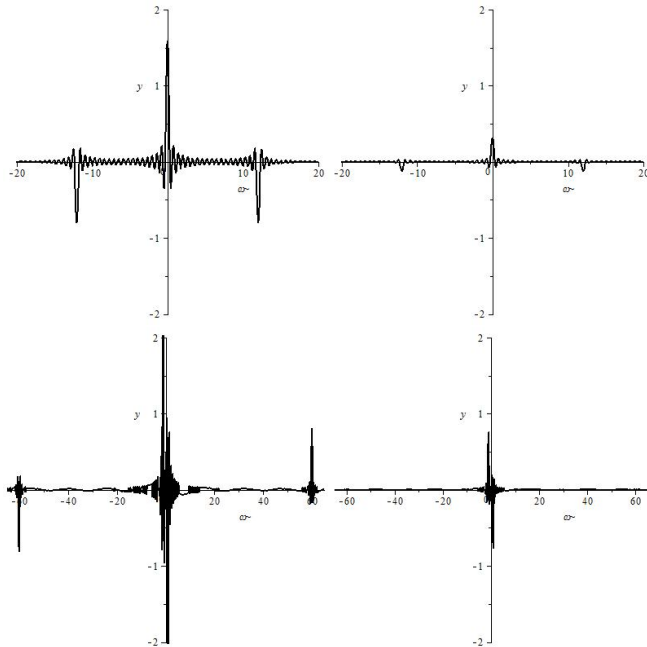
$$\mathcal{F}(1)(\Omega) = \sqrt{2\pi}\delta(\omega).$$

See 7.2.9 for comments on the *Dirac delta function*.  $\square$

**7.B.11.** Find the Fourier transform image of the convolution of the signals  $f(t)$  and  $g(t)$  from the Problem 7.B.1. Recall that  $f(t) = \sin(t) + 0.4 [\sin(6t)]^2 - 0.2 \sin(60t)$  and  $g$  is the characteristic function of the interval  $(-\varepsilon, \varepsilon)$ . Assume that the signal is nonzero only in the interval  $(-\Omega, \Omega)$ .

**Solution.** Once we note that  $\mathcal{F}(f * g) = \sqrt{2\pi}\mathcal{F}(f)\mathcal{F}(g)$ , we have all the ingredients ready. Indeed, in 7.B.5 and in the last two problems above, we already computed the Fourier image of  $g$  and of the sine and cosine functions on the interval  $(-\Omega, \Omega)$ .

Instead of writing the explicit formulae for the result, we display illustrations of the real components of  $\mathcal{F}(f)$  and  $\mathcal{F}(f * g)$  in the first line, and similarly the imaginary components in the second line, all with  $\Omega = 5, \varepsilon = 1/10$ .



The reader should compare the diagrams of  $f$  and  $f * g$  in 7.B.1 to see that the higher frequencies in  $f$  are effectively canceled by this convolution, as expected.  $\square$

Further, the definition of the Fourier transform yields

$$\begin{aligned} \langle f, g \rangle &= (\mathcal{F}^{-1} \mathcal{F}(f * \hat{g}))(0) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathcal{F}(f * \hat{g}) e^{ix\omega} d\omega |_{x=0} \\ &= \int_{-\infty}^{\infty} \mathcal{F}f(\omega) \mathcal{F}\hat{g}(\omega) d\omega \end{aligned}$$

while

$$\mathcal{F}(\hat{g})(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \overline{g(-x)} e^{-i\omega x} dx = \overline{\mathcal{F}(g)}(\omega).$$

Consequently,  $\langle f, g \rangle = \langle \mathcal{F}(f), \mathcal{F}(g) \rangle$ . Thus, we have verified that the Fourier transform preserves the scalar product and so it also preserves the  $L_2$ -norm.

This also explains our choice of the constants in the definition.

**7.2.9. Dirac delta-function.** We return to the first example of the inverse transform to the indicator function  $f_\Omega$  of the interval  $[-\Omega, \Omega]$ . Let  $\Omega$  approach infinity and denote by  $\sqrt{2\pi}\delta(t)$  the desired “limit function” for  $\mathcal{F}^{-1}(f_\Omega)(t)$ .

The inverse image of a product with an arbitrary image  $\mathcal{F}(g)$  can be expressed using convolution:

$$\mathcal{F}^{-1}(f_\Omega \cdot \mathcal{F}(g))(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(t) \mathcal{F}^{-1}(f_\Omega)(z - t) dt.$$

As  $\Omega$  increases to  $\infty$ , the left-hand expression should approach  $\mathcal{F}^{-1}(\mathcal{F}(g))(z) = g(z)$ , while on the right-hand side, we get

$$g(z) = \int_{-\infty}^{\infty} g(t) \delta(z - t) dt.$$

The desired  $\delta(t)$  thus looks as a “function” which takes zero everywhere except the single point  $t = 0$  where it “has an infinite value”. Integrating the product of  $\delta(t)$  with any integrable function  $g$  gives just the value of  $g$  at the point  $t = 0$ . Of course, this is strictly not a function at all. Nevertheless it is a useful concept. It is called the *Dirac function*  $\delta$  and it can be described correctly as an example of what is known as a distribution. Since we do not have enough space and time, we do not pay further attention to distributions. We mention only that the Dirac  $\delta$  can be imagined as a unit impulse at a single point. In fact, we saw similar concepts under the name “measure” when dealing with the Riemann-Stieltjes and Henstock-Kurzweil integrals, cf. 6.3.18, 6.3.14 and we shall come back to them in Chapter 10 in the context of probability. In this sense, the Dirac function is the (probability) measure concentrated in the origin and it can be realized by the Riemann-Stieltjes integral with the piecewise constant function  $g$  with the single unit jump in the origin. Its Fourier transform is the constant function  $\mathcal{F}(\delta)(\omega) = \frac{1}{\sqrt{2\pi}}$ .

On the other hand, many functions which are not strictly integrable on  $\mathbb{R}$  are Fourier-transformed to expressions with

As discussed in 7.2.5, the Fourier transform has an inverse operation. This means that no information is lost when changing from the time behaviour of a signal to its frequency behaviour. This allows us to use Fourier transform for the solution of functional equations involving differentiation or integration. We stay with elementary observations only and return to differential equations in one and more variables in the following chapters.



**7.B.12.** By using the inverse Fourier transform solve the integral equation

$$\int_0^\infty f(t) \sin(xt) dt = e^{-x}, \quad x > 0$$

for an unknown function  $f$ .

**Solution.** Multiply both sides of the equation by  $\sqrt{2/\pi}$ , to obtain the sine Fourier transform on the left-hand side. Apply the inverse transform to both sides of the equation to get

$$f(t) = \frac{2}{\pi} \int_0^\infty e^{-x} \sin(xt) dx, \quad t > 0.$$

Integrating by parts twice shows that

$$\int e^{-x} \sin(xt) dx = \frac{e^{-x}}{1+t^2} [-\sin(xt) - t \cos(xt)] + C.$$

Hence

$$\begin{aligned} & \int_0^\infty e^{-x} \sin(xt) dx \\ &= \lim_{x \rightarrow \infty} \left( \frac{e^{-x}}{1+t^2} [-\sin(xt) - t \cos(xt)] \right) - \frac{e^0(-t)}{1+t^2} = \frac{t}{1+t^2}. \end{aligned}$$

So

$$f(t) = \frac{2}{\pi} \frac{t}{1+t^2}, \quad t > 0.$$

□

**7.B.13.** Use the Fourier transform to find solutions of the non-homogeneous linear differential equation

$$(1) \quad y' = ay + f$$

where  $a \in \mathbb{R}$  is a non-zero constant and  $f$  is a known function. Can all solutions be obtained in this way?

**Solution.** The key observation for this problem is the relation between the Fourier transform and the derivative

$$\mathcal{F}(f')(\omega) = i\omega \mathcal{F}(f)(\omega),$$

see 7.2.6. Thus, if the Fourier transform is applied to the equation (1), we get the algebraic equation for  $\tilde{y} = \mathcal{F}(y)$

$$i\omega \tilde{y} = a\tilde{y} + \tilde{f}.$$

If it is assumed that  $\mathcal{F}(f) = \tilde{f}$  exists and there is a solution  $y$  with the Fourier image  $\tilde{y}$ , then

$$\tilde{y} = \frac{1}{i\omega - a} \tilde{f}$$

the Dirac  $\delta$ . For instance,

$$\mathcal{F}(\cos(nt))(\omega) = \sqrt{\frac{\pi}{2}} (\delta(n - \omega) + \delta(n + \omega)),$$

which can be seen from the calculation of the Fourier transform of the function  $f_\Omega \cos(nx)$  and then letting  $\Omega$  approach  $\infty$ , see the solution to problem 7.B.10.

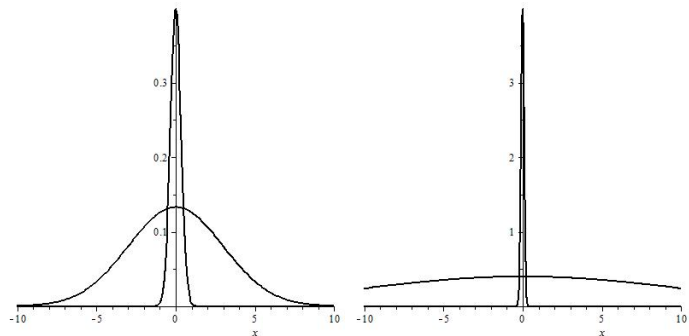
We can obtain the Fourier transform of the sine function in a similar way. We can take advantage of the fact that the transform of the derivative of this function differ only by a multiple of the imaginary unit and the new variable. Alternatively, we can also use the fact that the sine function is obtained from the cosine function by the phase shift of  $\pi/2$ .

These transforms are a basis for Fourier analysis of signals (see also problem 7.B.9): If a signal is a pure sinusoid of a given frequency, then this is recognized in the Fourier transform as two single-point impulses exactly at the positive and negative value of the frequency. If the signal is a linear combination of several such pure signals, then we obtain the same linear combination of single-point impulses. However, since we always process a signal in a finite time interval only, we get not single-point impulses, but rather a wavy curve similar to the function sinc with a strong maximum at the value of the corresponding frequency. The size of this maximum also yields information about the amplitude of the original signal.

Another good way how to approximate the Dirac delta function is to exploit the Gaussian functions. As seen in the solution to problem 7.B.7, the Fourier image of the Gaussian function

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}$$

is again Gaussian corresponding to the reciprocal values of  $\sigma$ . In the limit  $\sigma \rightarrow 0$ , the image converges fast to the multiple of constant function, see the illustrations on the picture, with  $\sigma = 3$  and  $\sigma = 1/10$ .



Notice that the rather large  $\sigma$  in the first illustration corresponds to a wide Gaussian, while the image is the slim one. The second illustration provides the opposite case. The preimage is the narrow Gaussian and the image is already reasonably close to the constant function. The Gaussians are chosen with  $L_1$ -norm equal to one, but the Fourier transform preserves the  $L_2$ -norm of the functions.

**7.2.10. Fourier sine and cosine transform.** If we apply the Fourier transform to an odd function  $f(t)$ , where  $f(-t) =$

and using the general relation  $\mathcal{F}^{-1}(g \cdot h) = \sqrt{2\pi}\mathcal{F}^{-1}f * \mathcal{F}^{-1}g$  between the product and convolutions from 7.2.7 we arrive at the final formula

$$y = \frac{1}{\sqrt{2\pi}}\mathcal{F}^{-1}\left(\frac{1}{i\omega - a}\right) * f.$$

So it is necessary to compute the inverse Fourier transform of the simple rational function  $(i\omega - a)^{-1}$ . Guess the solution in two steps.

Assume first  $a < 0$  and evaluate

$$\int_{-\infty}^0 e^{at} e^{-i\omega t} dt = \left[\frac{1}{a-i\omega} e^{(a-i\omega)t}\right]_{-\infty}^0 = \frac{1}{a-i\omega}.$$

Similarly for  $a > 0$

$$\int_0^{\infty} e^{at} e^{-i\omega t} dt = \left[\frac{1}{a-i\omega} e^{(a-i\omega)t}\right]_0^{\infty} = \frac{1}{i\omega - a}.$$

This provides the two desired results. Indeed, if the equation (1) comes with  $a > 0$ , we rewrite our rational function as  $-(a - i\omega)^{-1}$ . Next, the function  $-\sqrt{2\pi} e^{-at}$  for negative  $t$  provides the requested Fourier image. Immediately it is seen that the convolution

$$y(t) = -\int_{-\infty}^0 e^{ax} f(t-x) dx$$

is a solution. (The multiples  $\sqrt{2\pi}$  in the expression with the convolution cancel.) Similarly, if  $a < 0$  then

$$y(t) = \int_0^{\infty} e^{ax} f(t-x) dx$$

is a solution.

Not all solutions can be obtained in this way. For example,  $y' = y$  leads to  $y(t) = C e^t$  with an arbitrary constant  $C$ , but this is not a function with a Fourier image. With  $f(t) = 0$ , our procedure produces the zero function, which is just one of the solutions. Similarly, if we deal with the equation  $y' = y + t$ , then the particular solution suggested above is

$$y(t) = -\int_{-\infty}^0 e^x(t-x) dx = -t - 1. \quad \square$$

**7.B.14.** Check directly that the two functions  $y(t)$  found above are indeed solutions to the equation  $y' = ay + f$ .

○

**7.B.15.** As in the previous problem, solve the second order equation

$$y'' = ay + f.$$

**Solution.** Use the fact that  $\mathcal{F}(y'')(\omega) = -\omega^2\mathcal{F}(y)(\omega)$  and deduce the algebraic relation  $-\omega^2\tilde{y} = a\tilde{y} + \tilde{f}$ , for the Fourier images  $\tilde{y}$  and  $\tilde{f}$ . Hence

$$\tilde{y} = \frac{-1}{\omega^2 + a}\tilde{f}.$$

$-f(t)$ , the contribution in the integration of the product of  $f(t)$  and the function  $\cos(\pm\omega t)$  cancels for positive and negative values of  $t$ . Thus if  $f$  is odd, then

$$\mathcal{F}(f)(\omega) = \frac{-2i}{\sqrt{2\pi}} \int_0^{\infty} f(t) \sin \omega t dt.$$

The resulting function is odd again, hence for the same reason, the inverse transform can be determined similarly:

$$\mathcal{F}^{-1}(f)(\omega) = \frac{2i}{\sqrt{2\pi}} \int_0^{\infty} f(t) \sin \omega t dt.$$

Omitting the imaginary unit  $i$  (notice, this means we have to multiply the Fourier transform and its inverse by  $i$  and  $-i$ , respectively) inverse) gives mutually inverse transforms, which are called the *Fourier sine transform* for odd functions:

$$\tilde{f}_s(\omega) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \sin(\omega t) dt,$$

$$f(t) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} \tilde{f}_s(t) \sin(\omega t) dt.$$

Similarly, we can define the *Fourier cosine transform* for even functions:

$$\tilde{f}_c(\omega) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(t) \cos(\omega t) dt,$$

$$f(t) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} \tilde{f}_c(t) \cos \omega t dt.$$

**7.2.11. Laplace transforms.** The Fourier transform can be mainly applied to functions which are integrable in absolute value over  $\mathbb{R}$ . The *Laplace transform* is similar to the Fourier transform and does apply to all functions whose growth is not too fast:

$$\mathcal{L}(f)(s) = \int_0^{\infty} f(t) e^{-st} dt.$$

The integral operator  $\mathcal{L}$  has a rapidly reducing kernel if  $s$  is a positive real number. Therefore, the Laplace transform is usually perceived as a mapping of suitable functions on the interval  $[0, \infty)$  to the function on the same or shorter interval. The image  $\mathcal{L}(p)$  exists, for example, for every polynomial  $p(t)$  and all positive numbers  $s$ .

Analogously to the Fourier transform, we obtain the formula for the Laplace transform of a differentiated function for  $s > 0$  by using integration by parts:

$$\begin{aligned} \mathcal{L}(f'(t))(s) &= \int_0^{\infty} f'(t) e^{-st} dt \\ &= [f(t) e^{-st}]_0^{\infty} + s \int_0^{\infty} f(t) e^{-st} dt \\ &= -f(0) + s\mathcal{L}(f)(s). \end{aligned}$$

The properties of the Laplace transform and many other transforms used in technical practice can be found in specialized literature. We provide a few examples in the other column starting with 7.B.17.

In order to guess the correct preimage of the rational function in question, first assume  $a > 0$  and compute

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-a|x|} e^{-i\omega x} dx &= \left[ \frac{1}{a-i\omega} e^{a-i\omega x} \right]_{-\infty}^0 + \left[ \frac{-1}{a+i\omega} e^{-a-i\omega x} \right]_0^{\infty} \\ &= \frac{1}{a-i\omega} + \frac{1}{a+i\omega} = \frac{2a}{a^2 + \omega^2} \end{aligned}$$

Thus it is verified that

$$\mathcal{F}(e^{-\sqrt{a}|x|}) = \sqrt{\frac{a}{2\pi}} \frac{1}{a + \omega^2}.$$

Immediately (the factors  $\sqrt{2\pi}$  cancel)

$$y(t) = \frac{-1}{\sqrt{a}} e^{-\sqrt{a}|t|} * f(t) = \frac{-1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-\sqrt{a}|x|} f(t-x) dx.$$

The case  $a < 0$  is a little more complicated. But we may ask Maple or look up in the literature that the function

$$g(t) = \sin(b|t|)$$

has the Fourier image

$$\mathcal{F}(g)(\omega) = \frac{1}{2\pi} \frac{2b}{b^2 - \omega^2}.$$

We are nearly finished. The required preimage is

$$h(t) = \frac{\sqrt{2\pi}}{\sqrt{-4a}} \sin(\sqrt{-a}|t|)$$

and the resulting convolution is


$$y(t) = \frac{1}{\sqrt{-4a}} \int_{-\infty}^{\infty} \sin(\sqrt{-a}|x|) f(t-x) dx.$$

If we rewrite the equation as  $y'' + by = f$  with  $b > 0$ , the result says

$$y(t) = \frac{1}{2b} \int_{-\infty}^{\infty} \sin(b|x|) f(t-x) dx. \quad \square$$

**7.B.16.** Check directly that the two functions  $y(t)$  found above are indeed solutions to the equation  $y'' = ay + f$ .

○

 *The Laplace transfer is another integral transform which interchanges differentiation and algebraic multiplication. As with the Fourier transform, this is based on the properties of the exponential function, but this time we take the real exponential, see 7.2.11 for the formula. One advantage is that every polynomial has its Laplace image.*

**7.B.17.** Determine the Laplace transform  $\mathcal{L}(f)(s)$  for each of the functions

- (a)  $f(t) = e^{at}$ ;
- (b)  $f(t) = c_1 e^{a_1 t} + c_2 e^{a_2 t}$ ;
- (c)  $f(t) = \cos(bt)$ ;
- (d)  $f(t) = \sin(bt)$ ;
- (e)  $f(t) = \cosh(bt)$ ;
- (f)  $f(t) = \sinh(bt)$ ;
- (g)  $f(t) = t^k, k \in \mathbb{N}$ ,



**7.2.12. Discrete transforms.** Fourier analysis of signals mentioned in the previous paragraph are realized by special analog circuits in, for example, radio technology. Nowadays, we work only with discrete data when processing signals by computer circuits. Assume that there is a fixed (small) *sampling interval*  $\tau$  given in a (discrete) time variable and that, for a large natural number  $N$ , the signal repeats with period  $N\tau$ , which is the maximal period which can be represented in our discrete model.

We should not be surprised that our continuous models allow for a discrete analogy. Consider an  $N$ -dimensional vector, which can be imagined as the function  $r \mapsto f(r) \in \mathbb{C}$ , for  $r = 0, 1, \dots, N-1$ . Denote  $\Delta\omega = \frac{2\pi}{N}$  and  $\omega_k = k\Delta\omega$ . The simplest discrete approximation of the Fourier transform integral suggests that

$$\tilde{f}(k) = \frac{1}{N} \sum_{r=0}^{N-1} f(r) e^{-i\frac{2\pi}{N}kr}$$

should be a promising transformation  $f \mapsto \tilde{f}$ , whose inverse  $\tilde{f} \mapsto \hat{f}$  should not be far from

$$\hat{f}(k) = \sum_{r=0}^{N-1} \tilde{f}(r) e^{i\frac{2\pi}{N}kr}.$$

Actually, these are already the mutually inverse transformations:

**Theorem.** *The transformation above satisfies  $\hat{\tilde{f}}(k) = f(k)$  for all  $k = 0, 1, \dots, N-1$ .*

**PROOF.** Let

$$T = \sum_{r=0}^{N-1} e^{ir\frac{2\pi}{N}k}.$$

Then

$$e^{i\frac{2\pi}{N}k} T = \sum_{r=1}^N e^{ir\frac{2\pi}{N}k}$$

and so by subtraction,

$$(1 - e^{i\frac{2\pi}{N}k})T = 1 - e^{i2\pi k}.$$

The right hand side is 0 for all integers  $k$ . On the left side, the coefficient of  $T$  is not zero unless  $k$  is a multiple of  $N$ . Hence

$$T = \sum_{r=0}^{N-1} e^{ir\frac{2\pi}{N}k} = \begin{cases} N & \text{if } k \text{ is a multiple of } N \\ 0 & \text{otherwise.} \end{cases}$$

With  $k$  and  $s$  both confined to the range  $\{0, 1, 2, \dots, N-1\}$ ,  $k-s$  can only be a multiple of  $N$  when  $k = s$ . It follows that for such  $k$  and  $s$

$$\sum_{r=0}^{N-1} e^{ir\frac{2\pi}{N}(k-s)} = \delta_{ks} N$$

where the Kronecker delta  $\delta_{ks} = 0$  for  $k \neq s$  and  $\delta_{ks} = 1$  if  $k = s$ .

Finally, we compute:

where the constants  $b \in \mathbb{R}$  and  $a, a_1, a_2, c_1, c_2 \in \mathbb{C}$  are arbitrary. It is assumed that the positive number  $s \in \mathbb{R}$  is greater than the real parts of the numbers  $a, a_1, a_2 \in \mathbb{C}$ , and is greater than  $b$  in the problems (e) and (f).

**Solution.** The case (a). It follows directly from the definition of the Laplace transform that

$$\begin{aligned} \mathcal{L}(f)(s) &= \int_0^\infty e^{at} e^{-st} dt \\ &= \int_0^\infty e^{-(s-a)t} dt = \lim_{R \rightarrow \infty} \left( \frac{e^{-(s-a)R}}{-(s-a)} \right) - \frac{e^0}{-(s-a)} = \frac{1}{s-a}. \end{aligned}$$

The case (b). Using the result of the above case and the linearity of improper integrals, we obtain

$$\mathcal{L}(f)(s) = c_1 \int_0^\infty e^{a_1 t} e^{-st} dt + c_2 \int_0^\infty e^{a_2 t} e^{-st} dt = \frac{c_1}{s-a_1} + \frac{c_2}{s-a_2}.$$

The case (c). Since

$$\cos(bt) = \frac{1}{2} (e^{ibt} + e^{-ibt}),$$

the choice  $c_1 = 1/2 = c_2$ ,  $a_1 = ib$ ,  $a_2 = -ib$  in the previous case gives

$$\begin{aligned} \mathcal{L}(f)(s) &= \int_0^\infty \left( \frac{1}{2} e^{ibt} + \frac{1}{2} e^{-ibt} \right) e^{-st} dt = \\ &= \frac{1}{2(s-ib)} + \frac{1}{2(s+ib)} = \frac{s}{s^2+b^2}. \end{aligned}$$

The cases (d), (e), (f). Analogously, the choices

(d)  $c_1 = -i/2, c_2 = i/2, a_1 = ib, a_2 = -ib;$

(e)  $c_1 = 1/2 = c_2, a_1 = b, a_2 = -b;$

(f)  $c_1 = 1/2, c_2 = -1/2, a_1 = b, a_2 = -b$

lead to

(d)  $\mathcal{L}(f)(s) = \frac{b}{s^2+b^2};$

(e)  $\mathcal{L}(f)(s) = \frac{s}{s^2-b^2};$

(f)  $\mathcal{L}(f)(s) = \frac{b}{s^2-b^2}.$

Finally, the last one is obtained by a straightforward repetition of integration by parts:

$$\begin{aligned} \mathcal{L}(t^k)(s) &= \int_0^\infty t^k e^{-st} dt \\ &= \left[ -t^k \frac{1}{s} e^{-st} \right]_0^\infty + \frac{k}{s} \int_0^\infty t^{k-1} e^{-st} dt \\ &= \dots = \frac{k!}{s^k} \int_0^\infty e^{-st} dt = \frac{k!}{s^{k+1}}. \end{aligned}$$

**7.B.18.** Use the definition of the Gamma function  $\Gamma(t)$  in Chapter 6 in order to prove

$$\mathcal{L}(t^\alpha)(s) = \Gamma(\alpha + 1) \frac{1}{s^{\alpha+1}}$$

for general  $\alpha > 0$ . Compare the result to the one of 7.B.17(g).  $\circ$

**7.B.19.** For  $s > -1$ , calculate the Laplace transform  $\mathcal{L}(g)(s)$  of the function

$$g(t) = t e^{-t}.$$

Further, for  $s > 1$ , calculate the Laplace transform  $\mathcal{L}(h)(s)$  of the function

$$h(t) = t \sinh t.$$

$$\begin{aligned} \hat{f}(k) &= \sum_{r=0}^{N-1} \left( \frac{1}{N} \sum_{s=0}^{N-1} f(s) e^{-i \frac{2\pi}{N} r s} \right) e^{i \frac{2\pi}{N} r k} \\ &= \frac{1}{N} \sum_{s=0}^{N-1} f(s) \left( \sum_{r=0}^{N-1} e^{i \frac{2\pi}{N} r (k-s)} \right) \\ &= \frac{1}{N} \sum_{s=0}^{N-1} f(s) \delta_{ks} N = f(k). \end{aligned}$$

□

The computations in the proof also verify that the Fourier image of a periodic complex valued function with a unique period among the chosen sampling periods is just its amplitude at this particular frequency. Thus, if the signal has been created as a superposition of periodic signals with the sampling frequencies only, we obtain the absolutely optimal result. However, if the transformed signal has a frequency not exactly available among the sampling frequencies, there are nonzero amplitudes at all the sampling frequencies in the Fourier image. This is called frequency leaking in the technical literature. There is a vast amount of literature devoted to fast implementation and exploitation of the discrete Fourier transform, as well as other similar discrete tools. This is an extremely active area of current research.

### 3. Metric spaces

At the end of the chapter, we will focus on the concepts of distance and convergence in a more abstract way. This also provides the conceptual background for some of the already derived properties of Fourier series and Fourier transform. We need these concepts in miscellaneous contexts later.

It is hoped that the subsequent pages are a useful (and hopefully manageable) trip into the world of mathematics for the competent or courageous!

**7.3.1. Metrics and norms.** When we discussed Fourier series, the distance between functions on a space of functions, was commonly referred to. Now we examine the concept of distance more thoroughly.

We are familiar with the distance of points  $x, y$  in the Euclidean space  $\mathbb{R}^n$  given by the size of the vector  $x - y$ .

A very different example of distance is the so called *discrete metric* defined on any set  $X$  as follows. Each element  $x \in X$  has got distance zero from itself and one from all other elements in  $X$ . Notice that the triangle inequality is strict here – making detours when walking from one element to another always increases the distance travelled.

Both of these distances are generalized in the following concept:



**7.B.20.** The basic Laplace transforms are enumerated in the following table:

$y(t)$	$\mathcal{L}(y)(s)$
$t^k$	$\frac{k!}{s^{k+1}}$
$e^{at}$	$\frac{1}{s-a}$
$t e^{at}$	$\frac{1}{(s-a)^2}$
$t^n e^{at}$	$\frac{n!}{(s-a)^{n+1}}$
$\sin \omega t$	$\frac{\omega}{s^2 + \omega^2}$
$\cos \omega t$	$\frac{s}{s^2 + \omega^2}$
$e^{at} \sin \omega t$	$\frac{\omega}{(s-a)^2 + \omega^2}$
$e^{at} (\cos \omega t + \frac{a}{\omega} \sin \omega t)$	$\frac{s}{(s-a)^2 + \omega^2}$
$t \sin \omega t$	$\frac{2\omega s}{(s^2 + \omega^2)^2}$
$\sin \omega t - \omega t \cos \omega t$	$\frac{2\omega^3}{(s^2 + \omega^2)^2}$

Establish the 5th and 6th rows of the table above using Euler's formula  $e^{i\omega t} = \cos \omega t + i \sin \omega t$ .

As expected, using the features of the Laplace transform allows us to find explicit solutions to some differential equations. By 7.D.18, it is straightforward to incorporate the initial conditions into the solution. We present just two such examples in the problems at the conclusion of this Chapter, see 7.D.21. We return to this topic in Chapter 8.

### C. Metric spaces

The concept of metric is an abstract version of what we understand as the distance in Euclidean geometry. It is always based on the triangle inequality. The axioms in Definition 7.3.1 follow the Euclidean experience, saying that our "distance" of two elements has to be strictly positive (except if the two elements coincide), should be symmetric in the arguments, and should satisfy the triangle inequality.

Other concepts available in the literature are more abstract and might lead to more general objects (the most important ones being pseudometrics, ultrametrics, and semi-metrics).<sup>3</sup>

**7.C.1.** The discrete metric space  $X$  is defined as the set  $X$  with the function  $d : X \times X \rightarrow \mathbb{R}$

$$d(x, y) = \begin{cases} 1 & x \neq y \\ 0 & x = y. \end{cases}$$

Show that this is a metric space according to Definition 7.3.1.

Show how to introduce a metric on Cartesian products of metric spaces, so that product of two discrete metric spaces is again discrete.

<sup>3</sup>The first axiomatic definition of a "traditional" metric was given by Maurice Fréchet in 1906. However, the name of the metric comes from Felix Hausdorff, who used this word in his work from 1914.

### AXIOMS OF A METRIC AND A NORM

A set  $X$  together with a mapping  $d : X \times X \rightarrow \mathbb{R}$  such that for all  $x, y, z \in X$ , the following conditions are satisfied

- (1)  $d(x, y) \geq 0$ ; and  $d(x, y) = 0$  if and only if  $x = y$ ,
- (2)  $d(x, y) = d(y, x)$ ,
- (3)  $d(x, z) \leq d(x, y) + d(y, z)$ ,

is called a *metric space*. The mapping  $d$  is a *metric* on  $X$ .

If  $X$  is a vector space over  $\mathbb{R}$  and  $\| \cdot \| : X \rightarrow \mathbb{R}$  is a function satisfying

- (4)  $\|x\| \geq 0$ ; and  $\|x\| = 0$  if and only if  $x = 0$ ,
- (5)  $\|\lambda x\| = |\lambda| \|x\|$ , for all scalars  $\lambda$ ,
- (6)  $\|x + y\| \leq \|x\| + \|y\|$ ,

then the function  $\| \cdot \|$  is called a *norm* on  $X$ , and the space  $X$  is then a *normed vector space*. Every norm defines a metric by setting  $d(x, y) = \|x - y\|$ .

The Euclidean distance in the vector spaces  $\mathbb{R}^n$  satisfies the above three requirements and it is given by a norm.

The  $L_1$  norm  $\| \cdot \|_1$  and the  $L_2$  norm  $\| \cdot \|_2$  on functions, satisfy the norm properties and thus define a metric on the relevant spaces of functions.

Of course, not every metric can be defined by a norm in this way. The discrete metric mentioned above is an example.

Metrics given by a norm have very specific properties since their behaviour on the whole space  $X$  can be derived from the properties in an arbitrarily small neighbourhood of the zero element  $x = 0 \in X$ .

**7.3.2. Convergence.** The concepts of (close) neighbourhoods of particular elements, convergence of sequences of elements and the corresponding "topological" concepts can be defined on abstract metric spaces in much the same way as in the case of the real and complex numbers and their sequences. See the beginning of the fifth chapter, 5.2.3–5.2.8. We can almost copy these paragraphs; although the proof of the theorem 5.2.8 is much harder. We begin with the concept of convergent sequences in a metric space  $X$  with metric  $d$ :

#### CAUCHY SEQUENCES

Consider an arbitrary sequence of elements  $x_0, x_1, \dots$  in  $X$ . Suppose that for any fixed positive real number  $\varepsilon$ ,

$$d(x_i, x_j) < \varepsilon$$

for all but finitely many pairs of terms  $x_i, x_j$  of the sequence.

In other words, for any given  $\varepsilon > 0$ , there is an index  $N$  such that the above inequality holds for all  $i, j > N$ . Loosely put, the elements of the sequence are eventually arbitrarily close to each other.

Such a sequence is called a *Cauchy sequence*.

Just as in the case of the real or complex numbers, we would like every Cauchy sequence of terms  $x_i \in X$  to converge to some  $x$  in the following sense:



**Solution.** All three axioms of a metric from 7.3.1 are obviously satisfied in our definition of the discrete metric space.

Consider two metric spaces  $X$  and  $Y$  with metrics  $d_X$  and  $d_Y$ . The first obvious idea seems to add the distances of the components, i.e.

$$d((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2).$$

Clearly this is a metric (verify in detail!), but if the metric spaces  $X$  and  $Y$  are discrete, then considering points  $u = (x_1, y_1)$  and  $w = (x_2, y_2)$  such that  $x_1 \neq x_2, y_1 \neq y_2$  we arrive at  $d(u, w) = 2$ . Thus, this is not a discrete metric space.

But there is another simple possibility of introducing a metric on  $X \times Y$  using the maximum of the distances:

$$d((x_1, y_1), (x_2, y_2)) = \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}.$$

We call this the *product of the metric spaces  $X$  and  $Y$* . The triangle inequality as well as the other axioms are obvious (write down the explicit arguments!). Moreover, if both  $X$  and  $Y$  are discrete, then  $d$  is also a discrete metric.  $\square$

**7.C.2.** Decide whether or not the following sets and mappings form a metric space:

- i)  $\mathbb{N}, d(m, n) = \gcd(m, n)$
- ii)  $\mathbb{N}, d(m, n) = \frac{\max(m, n)}{\gcd(m, n)} - 1$
- iii) World population,  $d(P_1, P_2) = n$ ,  
 $P_1 = X_0, X_1, \dots, X_{n+1} = P_2$  is the shortest sequence of people, such that  $X_i$  knows  $X_{i+1}$  for  $i = 0, \dots, n$ .

**Solution.**

- i) No. The "distance"  $d$  does not satisfy that  $d(m, m) = 0$ .
- ii) No. The first and second conditions in the definition 7.3.1 are fulfilled, but the triangle inequality (property (3)) is not. The distance of 8 and 9 is 8, the distance of 8 and 6 is 3 and the distance of 6 and 9 is 2, thus  $d(8, 9) > d(8, 6) + d(6, 9)$ .
- iii) No. The "distance" is not symmetric. It would be a metric space, if the definition the word "knows" is changed to mean "know each other".  $\square$

**7.C.3.** Consider the set of binary words of the length  $n$ . Define the distance between two words as the number of bits in which they differ. This is called the Hamming distance, see 12.5.2). Show that it defines a metric.

**Solution.** The first two axioms of a metric are satisfied. For the third one let the words  $x$  and  $z$  differ in  $k$  bits. Let  $y$  be another word. Then consider just  $k$  bits, in which  $x$  and  $z$  differ. Clearly  $y$  differs in each of these bits exactly from one each of  $x$  and  $z$ . Thus considering only the parts of words  $x_p, y_p, z_p$  in the  $k$  bits, we have  $d(x_p, y_p) + d(y_p, z_p) = d(x_p, z_p)$ . In the other bits, the words  $x$  and  $z$  are the same, while  $x$  and  $y$  or  $y$  and  $z$  may differ. Thus  $d(x, y) + d(y, z) \geq d(x, z)$  and the third axiom is satisfied also.  $\square$

CONVERGENT SEQUENCES

Let  $x_0, x_1, \dots$  be a sequence in a metric space  $X$  and let  $x$  be an element of  $X$ . We say that the sequence  $\{x_i\}$  *converges* to the element  $x$ , if, for every positive real number  $\varepsilon$ , there is an integer  $N > 0$ , so that  $i > N$  implies  $d(x_i, x) < \varepsilon$ .

By the triangle inequality, it follows that for each pair of terms  $x_i, x_j$  from a convergent sequence with sufficiently large indices,

$$d(x_i, x_j) \leq d(x_i, x) + d(x, x_j) < 2\varepsilon.$$

Therefore, every convergent sequence is a Cauchy sequence. Conversely however, not every Cauchy sequence is convergent. Metric spaces where every Cauchy sequence is convergent are called *complete metric spaces*.

**7.3.3. Topology, convergence, and continuity.** Just as in the case of the real numbers, we can formulate the convergence in terms of "open neighbourhoods".

OPEN AND CLOSED SETS

**Definition** The *open  $\varepsilon$ -neighbourhood* of an element  $x$  in a metric space  $X$  (or just  $\varepsilon$ -neighbourhood for short) is the set

$$\mathcal{O}_\varepsilon(x) = \{y \in X; d(x, y) < \varepsilon\}.$$

A subset  $U \subset X$  is *open* if and only if for all  $x \in U$ ,  $U$  contains some  $\varepsilon$ -neighbourhood of  $x$ .

We define a subset  $W \subset X$  to be *closed* if and only if its complement  $X \setminus W$  is an open set.

Instead of an  $\varepsilon$ -neighbourhood, we also talk about (open)  $\varepsilon$ -ball centered at  $x$ . In the case of a normed space, we can consider  $\varepsilon$ -balls centered at zero: along with  $x$ ,  $\varepsilon$ -balls determine an  $\varepsilon$ -neighbourhood.

The *limit points of a subset  $A \subset X$*  are defined as those elements  $x \in X$  such that there is a sequence of points in  $A$  other than  $x$  converging to  $x$ .

**Lemma.** A subset in a metric space is closed if and only if it contains all of its limit points

**PROOF.** Suppose  $A$  is closed and  $x$  is a limit point of  $A$  but not belonging to  $A$ . Then  $x \in X \setminus A$  which is open, so there is an  $\varepsilon$ -neighbourhood of  $x$  not intersecting with  $A$ . But in every  $\varepsilon$ -neighbourhood of  $x$ , there are infinitely many points of the set  $A$ , since  $x$  is a limit point. This is a contradiction.

Conversely, suppose  $A$  contains all of its limit points and suppose  $x \in X \setminus A$ . If in every  $\varepsilon$ -neighbourhood of the point  $x$ , there is a point  $x_\varepsilon \in A$ , then the choices  $\varepsilon = 1/n$  provide a sequence of points  $x_n \in A$  converging to  $x$ . But then, the point  $x$  would have to be a limit point, thus lying in  $A$ , which again leads to a contradiction.  $\square$

For every subset  $A$  in a metric space  $X$ , we define its *interior* as the set of those points  $x$  in  $A$  for which a neighbourhood of  $x$  also belongs to  $A$ . We define the closure  $\bar{A}$  of

**7.C.4.** Consider any connected subset  $S \subset \mathbb{R}^n$  (any two points in  $S$  can be connected with a path lying in  $S$ ). Define the distance between two points as the length of the shortest path between the points. Is it a metric on  $S$ ?

**Solution.** It is a metric. All the axioms of the metric are trivially satisfied. But this metric has a special significance. The principle of "shortest way" is often met in reality. Recall for example Fermat's principle (see 5.E.123) of the least time, where we measure the length of a path by the time it is traveled by light. Generally, shortest paths in a metric space are called *geodesics*.  $\square$

**7.C.5.** Consider a space of integrable function on the interval  $[a, b]$ . Define the ( $L_1$ ) distance of the functions  $f, g$  as

$$\|f, g\| = \int_a^b |f(x) - g(x)| dx$$

. Why it is not a metric space?

**Solution.** The first axiom of the metric space in 7.3.1 is not satisfied. Any function of zero measure has distance 0 from the null function.

But if we consider an equivalence where two functions are equivalent, if they differ by a function of measure zero, then we get the space  $S^0(a, b)$ . The given distance considered on the equivalence classes of this equivalence is the  $L_1$  metric.  $\square$

**7.C.6.** Let  $r$  be a rational number and  $p$  a prime number. Then  $r$  can be uniquely written in the form  $r = p^k \frac{u}{v}$ , where  $u \in \mathbb{Z}$  and  $v \in \mathbb{N}$  are coprime and  $p$  does not divide both, the numerator  $u$  and the denominator  $v$ . Consider the map  $\|\cdot\|_p : \mathbb{Q} \rightarrow \mathbb{R}, \|r\| \mapsto p^{-k}$ . Show that it is a norm on  $\mathbb{Q}$  as a vector space over  $\mathbb{Q}$ . It is called the *p-adic norm*.  $\circ$

**Solution.** It is an exercise in elementary number theory.  $\square$

**7.C.7.** Consider the power set (the set of all subsets) of a given finite set. Determine whether the functions  $d_1$  and  $d_2$ , defined for all subsets  $X, Y$  by



(a)  $d_1(X, Y) := |(X \cup Y) \setminus (X \cap Y)|$ ,

(b)  $d_2(X, Y) := \frac{|(X \cup Y) \setminus (X \cap Y)|}{|X \cup Y|}$ , for  $X \cup Y \neq \emptyset$ ,  
 $d_2(\emptyset, \emptyset) := 0$

are metrics. ( $|X|$  is meant the number of elements of a set  $X$ , thus the metric  $d_1$  measures the size of the symmetric difference of the sets, while  $d_2$  measures the relative symmetric difference.)

**Solution.** We omit verifications of the first and second conditions from the definition of a metric in exercises on deciding whether a particular mapping is a metric. We analyze the triangle inequality only.

The case (a). For any sets  $X, Y, Z$ ,

$$(1) \quad (X \cup Z) \setminus (X \cap Z) \subseteq ((X \cup Y) \setminus (X \cap Y)) \cup ((Y \cup Z) \setminus (Y \cap Z))$$

To show this, suppose first that  $x$  is an element satisfying  $x \in X$  and  $x \notin Z$ . Then either  $x \in Y$  in which case  $x \in (Y \cup Z) \setminus (Y \cap Z)$ , or  $x \notin Y$  in which case  $x \in (X \cup Y) \setminus (X \cap Y)$ .

a set  $A$  as the union of the original set  $A$  with the set of all limit points of  $A$ .

As easily as in the case of the real numbers, we can verify that the intersection of any system of closed sets as well as the union of any finite system of closed sets is also closed.

On the other hand, any union of open sets is again an open set. A finite intersection of open sets is again an open set. Prove these propositions by yourselves in detail!

We also advise the reader to verify that the interior of a set  $A$  equals the union of all open sets contained in  $A$ , (alternatively put, the interior of  $A$  is the largest open subset of  $A$ ). The closure of  $A$  is the intersection of all closed sets which contain  $A$ , (alternatively put, the closure of  $A$  is the smallest closed superset of  $A$ ).

The closed and open sets are the essential concepts of the mathematical discipline called *topology*. Without pursuing these ideas further, we have just familiarised ourselves with the *topology of the metric spaces*.

The concept of convergence can be reformulated now as follows. A sequence of elements  $x_i, i = 0, 1, \dots$ , in a metric space  $X$  converges to  $x \in X$  if and only if for every open set  $U$  containing  $x$ , all but finitely many points of our sequence lie in  $U$ .

Just as in the case of the real numbers, we can define *continuous mappings* between metric spaces:

CONTINUOUS MAPPINGS

Let  $W$  and  $Z$  be metric spaces. A mapping  $f : W \rightarrow Z$  is continuous if and only if the inverse image  $f^{-1}(V)$  of every open set  $V \subset Z$  is an open set in  $W$ . This is equivalent to the statement that  $f$  is continuous if and only if for every  $z = f(x) \in Z$  and positive real number  $\varepsilon$ , there is a positive real number  $\delta$  such that for all elements  $y \in W$  with distance  $d_W(x, y) < \delta, d_Z(z, f(y)) < \varepsilon$ .

Again, as in the case of real-valued functions, a mapping  $f$  from one metric space to another is continuous if and only if it preserves the convergence of sequences (check this yourselves!).

**7.3.4.  $L_p$ -norms.** Now we have the general tools with which we can look at examples of metric spaces created by finite-dimensional vectors or functions at our disposal. We restrict ourselves to an extraordinarily useful class of norms.

We begin with the real or complex finite-dimensional vector spaces  $\mathbb{R}^n$  and  $\mathbb{C}^n$ , and for a fixed real number  $p \geq 1$  and any vector  $z = (z_1, \dots, z_n)$ , we define

$$\|z\|_p = \left( \sum_{i=1}^n |z_i|^p \right)^{1/p}.$$

We prove that this indeed defines a norm. The first two properties from the definition are clear. It remains to prove the triangle inequality. For that purpose, we use *Hölder's inequality*:

It follows that  $x$  belongs to the union, that is, the right side of 1. By symmetry, the same result holds if  $x$  is an element satisfying  $x \notin X$  and  $x \in Z$ . Since then all possibilities when  $x$  belongs to the left side of 1 are accounted for, the inclusion 1 is established.

But then,

$$\begin{aligned} d_1(X, Z) &= |(X \cup Z) \setminus (X \cap Z)| \\ &\leq |((X \cup Y) \setminus (X \cap Y)) \cup ((Y \cup Z) \setminus (Y \cap Z))| \\ &\leq |(X \cup Y) \setminus (X \cap Y)| + |(Y \cup Z) \setminus (Y \cap Z)| \\ &= d_1(X, Y) + d_1(Y, Z). \end{aligned}$$

The case (b). Proceed similarly to the case of  $d_1$ . Denote by  $X'$  the complement of a set  $X$ . The equalities

$$\begin{aligned} (X \cup Y) \setminus (X \cap Y) &= \\ (X \cap Y' \cap Z) \cup (X \cap Y' \cap Z') \cup (X' \cap Y \cap Z) \cup (X' \cap Y \cap Z'), \\ (Y \cup Z) \setminus (Y \cap Z) &= \\ (X \cap Y \cap Z') \cup (X \cap Y' \cap Z) \cup (X' \cap Y \cap Z') \cup (X' \cap Y' \cap Z), \\ ((X \cup Z) \setminus (X \cap Z)) \cup (Y \setminus (X \cup Z)) &= (X \cap Y \cap Z') \cup \\ (X \cap Y' \cap Z') \cup (X' \cap Y \cap Z) \cup (X' \cap Y' \cap Z) \cup (X' \cap Y \cap Z'), \end{aligned}$$

which, again, can be proved by listing several possibilities, imply a stronger form of (1), namely

$$\begin{aligned} ((X \cup Z) \setminus (X \cap Z)) \cup (Y \setminus (X \cup Z)) &\subseteq \\ ((X \cup Y) \setminus (X \cap Y)) \cup ((Y \cup Z) \setminus (Y \cap Z)). \end{aligned}$$

Further, we invoke the inequality

$$\frac{|(X \cup Z) \setminus (X \cap Z)|}{|X \cup Z|} \leq \frac{|((X \cup Z) \setminus (X \cap Z)) \cup (Y \setminus (X \cup Z))|}{|X \cup Z \cup (Y \setminus (X \cup Z))|}, \quad X \cup Z \neq \emptyset.$$

This is based on calculations with non-negative numbers only since in general

$$\frac{x}{z} \leq \frac{x+y}{z+y}, \quad y \geq 0, z > 0, x \in [0, z].$$

Since

$$X \cup Z \cup (Y \setminus (X \cup Z)) = X \cup Y \cup Z,$$

we obtain

$$\begin{aligned} d_2(X, Z) &= \frac{|(X \cup Z) \setminus (X \cap Z)|}{|X \cup Z|} \\ &\leq \frac{|((X \cup Z) \setminus (X \cap Z)) \cup (Y \setminus (X \cup Z))|}{|X \cup Z \cup (Y \setminus (X \cup Z))|} \\ &\leq \frac{|((X \cup Y) \setminus (X \cap Y)) \cup ((Y \cup Z) \setminus (Y \cap Z))|}{|X \cup Y \cup Z|} \leq \\ &\frac{|(X \cup Y) \setminus (X \cap Y)| + |(Y \cup Z) \setminus (Y \cap Z)|}{|X \cup Y \cup Z|} \\ &\leq \frac{|(X \cup Y) \setminus (X \cap Y)|}{|X \cup Y|} + \frac{|(Y \cup Z) \setminus (Y \cap Z)|}{|Y \cup Z|} = \\ &d_2(X, Y) + d_2(Y, Z), \end{aligned}$$

if  $X \cup Z \neq \emptyset$  and  $Y \neq \emptyset$ . However, for  $X = Z = \emptyset$  or  $Y = \emptyset$ , the triangle inequality clearly still holds.

Therefore, both mappings are metrics. The metric  $d_1$  is quite elementary, but the other metric, the metric  $d_2$  has wider applications. In the literature, it is also known as *Jaccard's metric*.<sup>4</sup>  $\square$

<sup>4</sup>It is named after the biologist Paul Jaccard, who described the measure of similarity in insects populations using the function  $1 - d_2$  in 1908.

HÖLDER INEQUALITY

**Lemma.** For a fixed real number  $p > 1$  and every pair of  $n$ -tuples of non-negative real numbers  $x_i$  and  $y_i$ ,

$$\sum_{i=1}^n x_i y_i \leq \left( \sum_{i=1}^n x_i^p \right)^{1/p} \cdot \left( \sum_{i=1}^n y_i^q \right)^{1/q},$$

where  $1/q = 1 - 1/p$ .

**PROOF.** Denote by  $X$  and  $Y$  the expressions in the product on the right-hand side of the inequality to be proved. If all of the numbers  $x_i$  or all of the numbers  $y_i$  are zero, then the statement is clearly true. Therefore, we can assume that  $X \neq 0$  and  $Y \neq 0$ .

We need to use the fact that the exponential function is a convex function. This can be stated: the graph of the exponential function lies below any of its chords, and indeed, the second derivative is again the exponential function, thus it is always positive. Hence, for any  $a$  and  $b$ , with  $p, q$  as above,

$$e^{(1/p)a + (1/q)b} \leq (1/p)e^a + (1/q)e^b$$

(in fact, this is the trivial case of the Jensen inequality, see 6.D.18).

Now, for those  $k$  with  $x_k y_k \neq 0$  we define the numbers  $v_k$  and  $w_k$  so that

$$x_k = X e^{v_k/p}, \quad y_k = Y e^{w_k/q}.$$

Then

$$e^{v_k/p + w_k/q} \leq \frac{1}{p} e^{v_k} + \frac{1}{q} e^{w_k}.$$

By substitution, it follows immediately that

$$x_k y_k \leq XY \left( \frac{1}{p} \left( \frac{x_k}{X} \right)^p + \frac{1}{q} \left( \frac{y_k}{Y} \right)^q \right).$$

Summing over  $k = 1, \dots, n$ , gives (notice that adding the terms with  $x_k y_k = 0$  does not spoil the inequality)

$$\begin{aligned} \frac{1}{XY} \sum_{i=1}^n x_i y_i &\leq \frac{1}{pX^p} \sum_{i=1}^n x_i^p + \frac{1}{qY^q} \sum_{i=1}^n y_i^q \\ &= \frac{1}{pX^p} X^p + \frac{1}{qY^q} Y^q = \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Multiplying this inequality by  $XY$  finishes the proof.  $\square$

Now we can prove that  $\| \cdot \|_p$  is indeed a norm:

MINKOWSKI INEQUALITY

For every  $p > 1$  and all  $n$ -tuples of non-negative real numbers  $(x_1, \dots, x_n)$  and  $(y_1, \dots, y_n)$ ,

$$\left( \sum_{i=1}^n (x_i + y_i)^p \right)^{1/p} \leq \left( \sum_{i=1}^n x_i^p \right)^{1/p} + \left( \sum_{i=1}^n y_i^p \right)^{1/p}.$$

7.C.8. Let

$$d(x, y) := \frac{|x - y|}{1 + |x - y|}, \quad x, y \in \mathbb{R}.$$

Prove that  $d$  is a metric on  $\mathbb{R}$ .

**Solution.** We prove the triangle inequality only (the rest is clear). Introduce an auxiliary function

$$(1) \quad f(t) := \frac{t}{1+t}, \quad t \geq 0.$$

Note that  $f(s) - f(r) = \frac{s}{1+s} - \frac{r}{1+r} = \frac{s-r}{(1+s)(1+r)} > 0$ , whenever  $s > r \geq 0$ .

It follows that  $f$  is increasing, a fact which can also be verified by examining the first derivative. Therefore,

$$\begin{aligned} d(x, z) &= \frac{|x - z|}{1 + |x - z|} = \frac{|x - y + y - z|}{1 + |x - y + y - z|} \\ &\leq \frac{|x - y| + |y - z|}{1 + |x - y| + |y - z|} \\ &= \frac{|x - y|}{1 + |x - y| + |y - z|} + \frac{|y - z|}{1 + |x - y| + |y - z|} \\ &\leq \frac{|x - y|}{1 + |x - y|} + \frac{|y - z|}{1 + |y - z|} \\ &= d(x, y) + d(y, z), \quad x, y, z \in \mathbb{R}. \end{aligned}$$

□

The metrics in the next problems are defined by norms on vector spaces of functions. See the definitions and discussion in 7.3.1.

7.C.9. Determine the distance between the functions



$f(x) = x$ ,  $g(x) = -\frac{x}{\sqrt{1+x^2}}$ ,  $x \in [1, 2]$  as elements of the normed vector space  $\mathcal{S}^0[1, 2]$  of (piecewise) continuous functions on the interval  $[1, 2]$  with norm

- (a)  $\|f\|_1 = \int_1^2 |f(x)| \, dx$ ;  
 (b)  $\|f\|_\infty = \max\{|f(x)|; x \in [1, 2]\}$ .

**Solution.** The case (a). We need only compute the norm of the difference of the functions

$$\begin{aligned} \int_1^2 |f(x) - g(x)| \, dx &= \int_1^2 x + \frac{x}{\sqrt{1+x^2}} \, dx \\ &= \left(\frac{x^2}{2} + \sqrt{1+x^2}\right)\Big|_1^2 = \frac{3}{2} + \sqrt{5} - \sqrt{2}. \end{aligned}$$

The case (b). It is necessary to compute

$$\max_{x \in [1, 2]} |f(x) - g(x)| = \max_{x \in [1, 2]} \left(x + \frac{x}{\sqrt{1+x^2}}\right).$$

Since

$$\left(x + \frac{x}{\sqrt{1+x^2}}\right)' = 1 + \frac{1}{(\sqrt{1+x^2})^3} > 0, \quad x \in [1, 2],$$

it follows that  $f - g$  is increasing, and so attains its maximum at the right end point of the interval when  $x = 2$ . So

$$\max_{x \in [1, 2]} \left(x + \frac{x}{\sqrt{1+x^2}}\right) = 2 + \frac{2}{\sqrt{1+2^2}} = 2 + \frac{2}{\sqrt{5}}.$$

□

To verify this inequality, we can use the following observation.

$$\sum_{i=1}^n (x_i + y_i)^p = \sum_{i=1}^n x_i(x_i + y_i)^{p-1} + \sum_{i=1}^n y_i(x_i + y_i)^{p-1}$$

and by Hölder's inequality (recall  $p > 1$ ),

$$\begin{aligned} \sum_{i=1}^n x_i(x_i + y_i)^{p-1} &\leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} \left(\sum_{i=1}^n (x_i + y_i)^{(p-1)q}\right)^{1/q} \\ \sum_{i=1}^n y_i(x_i + y_i)^{p-1} &\leq \left(\sum_{i=1}^n y_i^p\right)^{1/p} \left(\sum_{i=1}^n (x_i + y_i)^{(p-1)q}\right)^{1/q}. \end{aligned}$$

Adding the last two inequalities, and taking into account that  $p + q = pq$ , and so  $(p - 1)q = pq - q = p$ , we arrive at

$$\frac{\sum_{i=1}^n (x_i + y_i)^p}{\left(\sum_{i=1}^n (x_i + y_i)^p\right)^{1/q}} \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} + \left(\sum_{i=1}^n y_i^p\right)^{1/p}.$$

that is

$$\left(\sum_{i=1}^n (x_i + y_i)^p\right)^{1-1/q} \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} + \left(\sum_{i=1}^n y_i^p\right)^{1/p},$$

or

$$\left(\sum_{i=1}^n (x_i + y_i)^p\right)^{1/p} \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} + \left(\sum_{i=1}^n y_i^p\right)^{1/p},$$

since  $1 - 1/q = 1/p$ . This is the *Minkowski inequality* which we wanted to prove.

Thus we have verified that on every finite-dimensional real or complex vector space, there is a class of norms  $\|\cdot\|_p$  for all  $p > 1$ . The case  $p = 1$  was considered earlier. We can also consider  $p = \infty$  by setting

$$\|z\|_\infty = \max\{|z_i|, i = 1, \dots, n\}.$$

This is obviously a norm.

We notice that Hölder's inequality can, in the context of these norms, be written for all  $x = (x_1, \dots, x_n)$ ,  $y = (y_1, \dots, y_n)$  as

$$\sum_{i=1}^n |x_i| \cdot |y_i| \leq \|x\|_p \cdot \|y\|_q$$

for all  $p \geq 1$  and  $q$  satisfying  $1/p + 1/q = 1$ . For  $p = 1$ , we set  $q = \infty$ .

**7.3.5.  $L_p$ -norms for sequences and functions.** Now we can easily define the  $L_p$ -norms on suitable infinite-dimensional vector spaces as well. We begin with sequences.

The vector space  $\ell_p$ ,  $p \geq 1$ , is the set of all sequences of real or complex numbers  $x_0, x_1, \dots$  such that

$$\sum_{i=0}^{\infty} |x_i|^p < \infty.$$

If  $x = (x_1, x_2, \dots) \in \ell_p$ ,  $p \geq 1$ , then the norm is given by

$$\|x\|_p = \left(\sum_{i=0}^{\infty} |x_i|^p\right)^{1/p}$$



The  $L_1$  or  $L_2$  distances, discussed in the beginning of this chapter (cf. 7.1.2), reflect the basic intuition about the distance between graphs of the functions. However, in practice we need to understand more subtle concepts of distances. The most obvious way is to include the derivatives in a way similar to the values of the functions.

**7.C.10.** Consider the space  $\mathcal{S}^1[a, b]$  of piecewise differentiable (real or complex) functions on the interval  $[a, b]$  and show that the formula

$$\|f\| = \left( \int_a^b |f(x)|^2 + \alpha^2 |f'(x)|^2 dx \right)^{1/2}$$

with any real  $\alpha \geq 0$  is a norm on this vector space (up to the identification of functions differing only in the points of discontinuity).

Compute the distance between functions  $f(x) = \sin(x) + 0.1[\sin(6x)]^2 - 0.03 \sin(60x)$  and  $g(x) = \sin(x)$  on the interval  $[-\pi, \pi]$  in this norm and explain its dependence on  $\alpha$ .

**Solution.** The formula

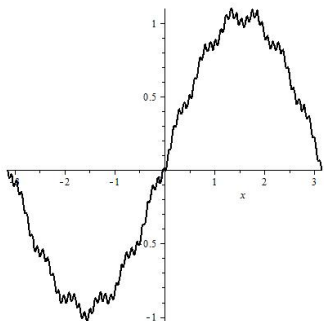
$$\langle f, g \rangle \mapsto \int_a^b f(x)\overline{g(x)} + \alpha^2 f'(x)\overline{g'(x)} dx$$

defines a scalar product on  $\mathcal{S}^1[a, b]$ . The mapping is linear in the first argument  $f$ , provides complex conjugate value if the arguments are exchanged and clearly is positive if  $f = g$  is non-zero on any interval (Ignore the values in the points of discontinuity, cf. the the discussion in 7.1.2). Thus the corresponding quadratic form defines a norm on the complex vector space  $\mathcal{S}^1[a, b]$ .

The distance in this norm is easily computed to obtain

$$\sqrt{0.02639 + 11.3097\alpha^2}.$$

Its dependence on  $\alpha$  can be seen in the illustration — the values of the function  $f(x)$  are nearly equal to  $\sin x$ , but the very wiggly difference is well apparent in the derivatives.



If  $\alpha = 0$ , the distance 0.162 is the usual  $L_2$  distance. If  $\alpha = 1$  the distance is 3.367.<sup>5</sup>  $\square$

<sup>5</sup>Here is an illustration of the very important concept of *Sobolev spaces*, where any number of derivatives can be involved. Moreover, we can use  $L_p$ ,  $p \geq 1$  in the definition of the norm instead of  $p = 2$ . There is much literature on this subject.

That  $\|x\|_p$  is a norm follows immediately from the Minkowski inequality by letting  $n \rightarrow \infty$ .

The vector space  $\ell_\infty$ , is the set of all bounded sequences of real or complex numbers  $x_0, x_1, \dots$

If  $x = (x_0, x_1, \dots) \in \ell_\infty$ , then its norm is given by

$$\|x\|_\infty = \sup\{|x_i|, i = 0, 1, 2, 3 \dots\}$$

It is easily checked that this is indeed a norm.

Eventually, we return to the space of functions  $\mathcal{S}^0[a, b]$  on a finite interval  $[a, b]$  or  $\mathcal{S}_c^0[a, b]$  on an unbounded interval. We have already met the  $L_1$  norm  $\|\cdot\|_1$ . However, for every  $p > 1$  and for all functions in such a space of functions, the Riemann integrals

$$\int_a^b |f(x)|^p dx$$

surely exist, so we can define

$$\|f\|_p = \left( \int_a^b |f(x)|^p dx \right)^{1/p}.$$

The Riemann integral was defined in terms of limits, using the Riemann sums which correspond to splittings  $\Xi$  with representatives  $\xi_i$ . In our case, those are the finite sums

$$S_{\Xi, \xi} = \sum_{i=1}^n |f(\xi_i)|^p (x_i - x_{i-1}).$$

Hölder's inequality applied to the Riemann sums of a product  $|f(x)g(x)|$  of two functions  $f(x)$  and  $g(x)$  gives

$$\begin{aligned} & \sum_{i=1}^n |f(\xi_i)||g(\xi_i)|(x_i - x_{i-1}) = \\ & = \sum_{i=1}^n |f(\xi_i)|(x_i - x_{i-1})^{1/p} |g(\xi_i)|(x_i - x_{i-1})^{1/q} \\ & \leq \left( \sum_{i=1}^n |f(\xi_i)|^p (x_i - x_{i-1}) \right)^{1/p} \left( \sum_{i=1}^n |g(\xi_i)|^q (x_i - x_{i-1}) \right)^{1/q}. \end{aligned}$$

where on the right-hand side, there is the product of the Riemann sums for the integrals  $\|f\|_p$  and  $\|g\|_q$ .

Moving to limits, we thus verify *Hölder's inequality for integrals*:

$$\int_a^b f(x)g(x) dx \leq \left( \int_a^b f(x)^p dx \right)^{1/p} \left( \int_a^b g(x)^q dx \right)^{1/q}$$

which is valid for all non-negative real-valued functions  $f$  and  $g$  in our space of piecewise continuous functions with compact support.

In just the same way as in the previous paragraph, we can derive the integral form of the Minkowski inequality from Hölder's inequality:

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Thus  $\|\cdot\|_p$  is indeed a norm on the vector space of all continuous functions having a compact support for all  $p > 1$  (we verified this for  $p = 1$  long ago).

Now we move to more theoretical considerations. Though these exercises may not look particularly practical, they should be of help in understanding the basic concepts of metric spaces, the convergence as well as their links to the topological concepts.



**7.C.11.** Show that the definition of a metric as a function  $d$  defined on  $X \times X$  for a non-empty set  $X$  and satisfying

- (1)  $d(x, y) = 0$ , if and only if  $x = y$ ,  $x, y \in X$ ,
- (2)  $d(x, z) \leq d(y, x) + d(y, z)$ ,  $x, y, z \in X$ ,

is equivalent to the definition given in the theoretical part, in paragraph 7.3.1.

**Solution.** At first glance, it seems that this definition demands fewer requirements on the metric than the definition from the theoretical part. The two definitions are equivalent if and only if the two conditions of non-degeneracy and triangle inequality imply

- (3)  $d(x, y) \geq 0$ ,  $x, y \in X$ ,
- (4)  $d(x, y) = d(y, x)$ ,  $x, y \in X$ .

However, if we set  $x = z$  in (2), we get the non-negativity of the metric from (1). Similarly, the choice  $y = z$  in (2) together with (1) implies that  $d(x, y) \leq d(y, x)$  for all points  $x, y \in X$ . Interchanging the variables  $x$  and  $y$  then gives  $d(y, x) \leq d(x, y)$ , i.e. (4). Thus, it is proved that the definitions are equivalent.  $\square$

**7.C.12.** Describe all sequences in a discrete metric space  $X$ , which are convergent or Cauchy.

**Solution.** Since the distance between two points  $x, y$  in  $X$  is either 1 or zero, the sequence  $x_1, x_2, \dots$  is Cauchy if and only if all  $x_i$  are equal, except for a finite number of them. But then, the sequence is convergent.  $\square$

This problem shows a behaviour quite different from the convergence of sequences in the metric spaces  $X = \mathbb{R}$  or  $X = \mathbb{C}$ . But sequences of integers would behave in a very similar way. On the other hand, we deal mostly with metrics on spaces of functions, where intuition gained in the real line  $\mathbb{R}$  may be useful.

**7.C.13.** Determine whether or not the sequence  $\{x_n\}_{n \in \mathbb{N}}$  where

$$x_1 = 1, \quad x_n = 1 + \frac{1}{2} + \dots + \frac{1}{n}, \quad n \in \mathbb{N} \setminus \{1\},$$

is a Cauchy sequence in  $\mathbb{R}$  using the standard metric.

**Solution.** Recall that

$$(1) \quad \sum_{k=1}^{\infty} \frac{1}{k} = \infty, \quad \text{i.e.} \quad \sum_{k=m}^{\infty} \frac{1}{k} = \infty, \quad m \in \mathbb{N}.$$

Therefore,

$$\lim_{n \rightarrow \infty} |x_n - x_m| = \sum_{k=m+1}^{\infty} \frac{1}{k} = \infty, \quad m \in \mathbb{N}.$$

We use the word “norm” for the entire space  $\mathcal{S}^0[a, b]$  of piecewise continuous functions in this context; however, we should bear in mind that we have to identify those functions which differ only by their values at points of discontinuity.

Among these norms, the case of  $p = 2$  is special because of the existence of the inner product. In this case, we could have derived the triangle inequality much more easily using the Schwarz inequality.

For the functions from  $\mathcal{S}^0[a, b]$ , we can define an analogy of the  $L_\infty$ -norm on  $n$ -dimensional vectors. Since our functions are piecewise continuous, they always have suprema of absolute values on a finite closed interval, so we can set

$$\|f\|_\infty = \sup\{f(x), x \in [a, b]\}$$

for such a function  $f$ . If we considered both the one-sided limits (which always exist by our definition) and the value of the function itself to be the value  $f(x)$  at points of discontinuity, we can work with maxima instead of suprema. It is apparent again that it is a norm (except for the problems with values at discontinuity points).

**7.3.6. Completion of metric spaces.** Both the real numbers



$\mathbb{R}$  and the complex numbers  $\mathbb{C}$  are (with the metric given by the absolute value) a complete metric space. This is contained in the axiom of the existence of suprema. Recall that the real numbers were created as a “completion” of the space of rational numbers which is not complete. It is evident that the closure of the set  $\mathbb{Q} \subset \mathbb{R}$  is  $\mathbb{R}$ .

#### DENSE AND NOWHERE-DENSE SUBSETS

We say that a subset  $A \subset X$  in a metric space  $X$  is *dense* if and only if the closure of  $A$  is the whole space  $X$ .

A set  $A$  is said to be *nowhere dense* in  $X$  if and only if the set  $X \setminus \bar{A}$  is dense.

Evidently,  $A$  is dense in  $X$  if every open set in the whole space  $X$  has a non-empty intersection with  $A$ .

In all cases of norms on functions from the previous paragraph, the metric spaces defined in this way are not complete since it can happen that the limit of a Cauchy sequence of functions from our vector space  $\mathcal{S}^0[a, b]$  should be a function which does not belong to this space any more. Consider the interval  $[0, 1]$  as the domain of functions  $f_n$  which take zero on  $[0, 1/n]$  and are equal to  $\sin(1/x)$  on  $[1/n, 1]$ . They converge to the function  $\sin(1/x)$  in all  $L_p$  norms, but this function does not lie in the space.

#### COMPLETION OF A METRIC SPACE

Let  $X$  be a metric space with metric  $d$  which is not complete. A metric space  $\tilde{X}$  with metric  $\tilde{d}$  such that  $X \subset \tilde{X}$ ,  $d$  is the restriction of  $\tilde{d}$  to the subset  $X$ , and the closure  $\bar{X}$  in  $\tilde{X}$  is the whole space  $\tilde{X}$  is called a *completion of the metric space*  $X$ .

Hence the sequence  $\{x_n\}$  is not a Cauchy sequence.

Alternatively,  $\{x_n\}$  is not a Cauchy sequence, since if it is, then it is convergent in the complete metric space  $\mathbb{R}$ , which contradicts the divergence shown in (1).  $\square$

**7.C.14.** Repeat the question from the previous problem with the metric  $d$  given by (cf. 7.C.8)

$$d(x, y) := \frac{|x - y|}{1 + |x - y|}, \quad x, y \in \mathbb{R}.$$

**Solution.** Instead of repeating the arguments, we point out the difference between the given metric from the standard one. The difference is expressed by the function  $f$  introduced in (1). This is a continuous function and, moreover a bijection between the sets  $[0, \infty)$  and  $[0, 1)$ , having the property that  $f(0) = 0$ . Further, the property of a sequence being Cauchy or convergent in a metric space is defined by being Cauchy or convergent for the real numbers describing the distances between the elements in the sequence. But the continuous mappings preserve convergence or the property being Cauchy, and hence the solution for the new metric is the same as with the standard one.  $\square$

**7.C.15.** Determine whether or not the metric space  $\mathcal{C}[-1, 1]$  of continuous functions on the interval  $[-1, 1]$  with metric given by the norm

- (a)  $\|f\|_p = \left( \int_{-1}^1 |f(x)|^p dx \right)^{1/p}$  for  $p \geq 1$ ;  
 (b)  $\|f\|_\infty = \max\{|f(x)|; x \in [-1, 1]\}$

is complete.

**Solution.** The case (a). For every  $n \in \mathbb{N}$ , define a function

$$f_n(x) = 0, \quad x \in [-1, 0), \quad f_n(x) = 1, \quad x \in \left(\frac{1}{n}, 1\right], \\ f_n(x) = nx, \quad x \in \left[0, \frac{1}{n}\right].$$

For every  $m \geq n, m, n \in \mathbb{N}$ , we compute the inequality

$$\left( \int_{-1}^1 |f_m(x) - f_n(x)|^p dx \right)^{1/p} < \left( \int_0^{1/n} 1 dx \right)^{1/p} = \left(\frac{1}{n}\right)^{1/p}.$$

It follows that the sequence  $\{f_n\}_{n \in \mathbb{N}} \subset \mathcal{C}[-1, 1]$  is a Cauchy sequence of functions.

Suppose the sequence  $\{f_n\}$  has a  $\|\cdot\|_p$  limit  $f$  in  $\mathcal{C}[-1, 1]$ . We show that this limit cannot be continuous in  $x = 0$ . For every  $\varepsilon \in (0, 1)$ , there exists an  $n(\varepsilon) \in \mathbb{N}$  such that

$$f_n(x) = 0, \quad x \in [-1, 0], \quad f_n(x) = 1, \quad x \in [\varepsilon, 1]$$

for all  $n \geq n(\varepsilon)$ . Imagine,  $f(y) \neq 1$  at some  $y \geq \varepsilon$ . Then  $\|f - f_n\| \geq \delta > 0$  for all  $n \geq n(\varepsilon)$  and some  $\delta$ , since  $f$  is continuous. Thus  $f \neq 1$  on some bounded interval containing  $y$ . Therefore  $f$  must satisfy

$$f(x) = 0, \quad x \in [-1, 0], \quad f(x) = 1, \quad x \in [\varepsilon, 1]$$

for an arbitrarily small  $\varepsilon > 0$ . Thus, necessarily,

$$f(x) = 0, \quad x \in [-1, 0], \quad f(x) = 1, \quad x \in (0, 1].$$

The following theorem says that the completion of an arbitrary (incomplete) metric space  $X$  can be found in essentially the same way as the real numbers were created from the rationals. (Actually, the reader might read the detailed proof below first having this completion of rationals in mind, which verifies that this construction leads to  $\mathbb{R}$  with the standard complete metric, thus satisfying the axioms of the reals.)

**7.3.7. Theorem.** Let  $X$  be a metric space with metric  $d$  which is not complete. Then there exists a completion  $\tilde{X}$  of  $X$ .

**PROOF.** The idea of the construction is identical to the one used when building the real numbers. Two Cauchy sequences  $x_i$  and  $y_i$  of points belonging to  $X$  are considered equivalent if and only if  $d(x_i, y_i)$  converges to zero for  $i$  approaching infinity. This is a convergence of real numbers, thus the definition is correct.

From the properties of convergence on the real numbers, it is clear that the relation defined above is an equivalence relation. The reader is advised to verify this in detail. For instance, the transitivity follows from the fact that the sum of two sequences converging to zero converges to zero as well.

We define  $\tilde{X}$  as the set of the classes of this equivalence of Cauchy sequences. The original points  $x \in X$  can be identified with the class of sequences equivalent to the constant sequence  $x_i = x, i = 0, 1, \dots$

It is now easy to define the metric  $\tilde{d}$ . We put

$$\tilde{d}(\tilde{x}, \tilde{y}) = \lim_{i \rightarrow \infty} d(x_i, y_i)$$

for sequences  $\tilde{x} = \{x_0, x_1, \dots\}$  and  $\tilde{y} = \{y_0, y_1, \dots\}$ .

First, we have to verify that this limit exists at all and is finite. Notice the consequence of the triangle inequality for differences. Taking  $x, y$ , and  $z \in X$ ,  $d(x, y) \leq d(x, z) + d(y, z)$  and thus,  $d(x, y) - d(x, z) \leq d(y, z)$ . Swapping  $y$  and  $z$ , we obtain  $d(x, z) - d(x, y) \leq d(y, z)$ . Thus,

$$(1) \quad |d(x, y) - d(x, z)| \leq d(y, z)$$

(draw a picture!). Now, the fact that both the sequences  $\tilde{x}$  and  $\tilde{y}$  are Cauchy sequences implies that the considered sequence is also a Cauchy sequence of real numbers:

$$|d(x_i, y_i) - d(x_j, y_j)| \leq |d(x_i, y_i) - d(x_i, y_j)| \\ + |d(x_i, y_j) - d(x_j, y_j)| \\ \leq d(y_i, y_j) + d(x_i, x_j).$$

Thus, the limit of  $d(x_i, y_i)$  exists.

If we select different representatives  $\tilde{x} = \{x'_0, x'_1, \dots\}$  and  $\tilde{y} = \{y'_0, y'_1, \dots\}$ , then by similar argument as above

$$|d(x'_i, y'_i) - d(x_i, y_i)| \leq |d(x'_i, y'_i) - d(x'_i, y_i)| \\ + |d(x'_i, y_i) - d(x_i, y_i)| \\ \leq d(x_i, x'_i) + d(y_i, y'_i).$$

Therefore, the definition is indeed independent of the choice of representatives.

We verify that  $\tilde{d}$  is a metric on  $\tilde{X}$ . The first and second properties are clear, so it remains to prove the triangle inequality. For that purpose, choose three Cauchy representatives of

But this function is not continuous on  $[-1, 1]$ , so it does not belong to the considered metric space. Therefore, the sequence  $\{f_n\}$  does not have a limit in  $\mathcal{C}[-1, 1]$ , so this space is not complete.

The case (b). Let an arbitrary Cauchy sequence  $\{f_n\}_{n \in \mathbb{N}} \subset \mathcal{C}[-1, 1]$  be given. The terms of this sequence are continuous functions  $f_n$  on  $[-1, 1]$  having the property that for  $\varepsilon > 0$  (or for every  $\varepsilon/2$  if you want) there is an  $n(\varepsilon) \in \mathbb{N}$  such that

$$(1) \quad \max_{x \in [-1, 1]} |f_m(x) - f_n(x)| < \frac{\varepsilon}{2}, \quad m, n \geq n(\varepsilon).$$

In particular, for every  $x \in [-1, 1]$ , we get a Cauchy sequence  $\{f_n(x)\}_{n \in \mathbb{N}} \subset \mathbb{R}$  of numbers. Since the metric space  $\mathbb{R}$  with the usual metric is complete, every (for  $x \in [-1, 1]$ ) sequence  $\{f_n(x)\}$  is convergent. Set

$$f(x) := \lim_{n \rightarrow \infty} f_n(x), \quad x \in [-1, 1].$$

Letting  $m \rightarrow \infty$  in (1), we obtain

$$\max_{x \in [-1, 1]} |f(x) - f_n(x)| \leq \frac{\varepsilon}{2} < \varepsilon, \quad n \geq n(\varepsilon).$$

It follows that the sequence  $\{f_n\}_{n \in \mathbb{N}}$  converges uniformly (that is, with respect to the given norm), to the function  $f$  on  $[-1, 1]$ . Since the uniform limit of continuous functions is continuous, so is  $f$ , so  $f \in \mathcal{C}[-1, 1]$ , see 6.3.4. Therefore, the metric space is complete.  $\square$

The same reasoning as above, and hence the same results, apply to the more general metric space  $\mathcal{C}[a, b]$  of continuous functions on any closed bounded interval  $[a, b]$  or on the space  $\mathcal{C}_c$  of continuous functions with compact support.

**7.C.16.** Prove that the metric space  $\ell_2$  is complete.

**Solution.** Recall that  $\ell_2$  is the space of sequences of real numbers with the  $L_2$ -norm, see 7.3.5.

Consider an arbitrary Cauchy sequence  $\{x_n\}_{n \in \mathbb{N}}$  in the space  $\ell_2$ . Every term of this sequence is again a sequence, i.e.,  $x_n = \{x_n^k\}_{k \in \mathbb{N}}$ ,  $n \in \mathbb{N}$ . Of course, the range of indices does not matter – there is no difference whether  $n, k \in \mathbb{N}$  or  $n, k \in \mathbb{N} \cup \{0\}$ . Introduce auxiliary sequences  $y_k$  for  $k \in \mathbb{N}$  so that

$$y_k = \{y_k^n\}_{n \in \mathbb{N}} = \{x_n^k\}_{n \in \mathbb{N}}.$$

If  $\{x_n\}$  is a Cauchy sequence in  $\ell_2$ , then each of the sequences  $y_k$  is a Cauchy sequence in  $\mathbb{R}$  (the sequences  $y_k$  are sequences of real numbers). It follows from the completeness of  $\mathbb{R}$  (with respect to the usual metric) that all of the sequences  $y_k$  are convergent. Denote their limits by  $z_k$ ,  $k \in \mathbb{N}$ .

It suffices to prove that  $z = \{z_k\}_{k \in \mathbb{N}} \in \ell_2$  and that the sequence  $\{x_n\}$  converges for  $n \rightarrow \infty$  in  $\ell_2$  just to the sequence  $z$ . The sequence  $\{x_n\}_{n \in \mathbb{N}} \subset \ell_2$  is a Cauchy sequence; therefore, for every  $\varepsilon > 0$ , there is an  $n(\varepsilon) \in \mathbb{N}$  with the property that

$$\sum_{k=1}^{\infty} (x_m^k - x_n^k)^2 < \varepsilon, \quad m, n \geq n(\varepsilon), \quad m, n \in \mathbb{N}.$$

In particular,

$$\sum_{k=1}^l (x_m^k - x_n^k)^2 < \varepsilon, \quad m, n \geq n(\varepsilon), \quad m, n, l \in \mathbb{N},$$

the elements  $\tilde{x}, \tilde{y}, \tilde{z}$ , and we obtain

$$\begin{aligned} \tilde{d}(\tilde{x}, \tilde{z}) &= \lim_{i \rightarrow \infty} d(x_i, z_i) \\ &\leq \lim_{i \rightarrow \infty} d(x_i, y_i) + \lim_{i \rightarrow \infty} d(y_i, z_i) \\ &= \tilde{d}(\tilde{x}, \tilde{y}) + \tilde{d}(\tilde{y}, \tilde{z}). \end{aligned}$$

The restriction of the metric  $\tilde{d}$  just defined to the original space  $X$  is identical to the original metric because the original points are represented by constant sequences.

It is required to prove that  $X$  is dense in  $\tilde{X}$ . Let  $\tilde{x} = \{x_i\}$  be a fixed Cauchy sequence, and let  $\varepsilon > 0$  be given. Since the sequence  $x_i$  is a Cauchy sequence, all pairs of its terms  $x_n, x_m$  for sufficiently large indices  $m$  and  $n$  become closer to each other than  $\varepsilon$ . Then the choice  $y = x_n$  for one of those indices necessarily implies that the elements  $y$  and  $x_m$  are closer together than  $\varepsilon$ , and so,  $\tilde{d}(\tilde{y}, \tilde{x}) \leq \varepsilon$ . Hence there is an element  $y$  of the original space such that the distance of the sequences of  $y$ 's from the chosen sequence  $x_i$  does not exceed  $\varepsilon$ . This establishes the denseness of  $X$ .

It remains to prove that the constructed metric space is complete. That is, that Cauchy sequences of points of the extended space  $\tilde{X}$  with respect to the metric  $\tilde{d}$  are necessarily convergent to a point in  $\tilde{X}$ . This can be done by approximating the points of a Cauchy sequence  $\tilde{x}_k$  by points  $y_k$  from the original space  $X$  so that the resulting sequence  $\tilde{y} = \{y_i\}$  would be the limit of the original sequence with respect to the metric  $\tilde{d}$ .

Since  $X$  is a dense subset in  $\tilde{X}$ , we can choose, for every element  $\tilde{x}_k$  of our fixed sequence, an element  $z_k \in X$  so that the constant sequence  $\tilde{z}_k$  would satisfy  $\tilde{d}(\tilde{x}_k, \tilde{z}_k) < 1/k$ . Now consider the sequence  $\tilde{z} = \{z_0, z_1, \dots\}$ . The original sequence  $\tilde{x}$  is Cauchy. So for a fixed real number  $\varepsilon > 0$ , there is an index  $n(\varepsilon)$  such that  $\tilde{d}(\tilde{x}_n, \tilde{x}_m) < \varepsilon/2$  whenever both  $m$  and  $n$  are greater than  $n(\varepsilon)$ . Without loss of generality, the index  $n(\varepsilon)$  is greater than or equal to  $4/\varepsilon$ . Now, for  $m$  and  $n$  greater than  $n(\varepsilon)$ , we get:

$$\begin{aligned} d(z_m, z_n) &= \tilde{d}(\tilde{z}_m, \tilde{z}_n) \\ &\leq \tilde{d}(\tilde{z}_m, \tilde{x}_m) + \tilde{d}(\tilde{x}_m, \tilde{x}_n) + \tilde{d}(\tilde{x}_n, \tilde{z}_n) \\ &\leq 1/m + \varepsilon/2 + 1/n \leq 2\frac{\varepsilon}{4} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

Hence it is a Cauchy sequence  $z_i$  of elements in  $X$ , and so  $\tilde{z} \in \tilde{X}$ . From the triangle inequality,

$$\tilde{d}(\tilde{z}, \tilde{x}_n) \leq \tilde{d}(\tilde{z}, \tilde{z}_n) + \tilde{d}(\tilde{z}_n, \tilde{x}_n).$$

From the previous bounds, both terms on the right-hand side converge to zero. Hence the distances  $\tilde{d}(\tilde{x}_n, \tilde{z})$  approach zero, thereby finishing the proof.  $\square$

**7.3.8. Uniqueness.** We consider now the uniqueness of the completion of metric spaces.



whence, letting  $m \rightarrow \infty$ ,

$$\sum_{k=1}^l (z_k - x_n^k)^2 \leq \varepsilon, \quad n \geq n(\varepsilon), \quad n, l \in \mathbb{N},$$

i.e. (this time  $l \rightarrow \infty$ )

$$(1) \quad \sum_{k=1}^{\infty} (z_k - x_n^k)^2 \leq \varepsilon, \quad n \geq n(\varepsilon), \quad n \in \mathbb{N}.$$

Especially,

$$\sum_{k=1}^{\infty} (z_k - x_n^k)^2 < \infty, \quad n \geq n(\varepsilon), \quad n \in \mathbb{N}$$

and, at the same time,

$$\sum_{k=1}^{\infty} (x_n^k)^2 < \infty, \quad n \in \mathbb{N},$$

which follows straight from  $\{x_n\}_{n \in \mathbb{N}} \subset \ell_2$ .

Since (cf. the special case of Hölder's inequality for  $p = 2$  in 7.3.4)

$$\sum_{k=1}^{\infty} (z_k x_n^k) \leq \sqrt{\sum_{k=1}^{\infty} z_k^2} \cdot \sqrt{\sum_{k=1}^{\infty} (x_n^k)^2}, \quad n \in \mathbb{N}$$

and

$$\sum_{k=1}^{\infty} (z_k - x_n^k)^2 = \sum_{k=1}^{\infty} (z_k^2 - 2z_k x_n^k + (x_n^k)^2), \quad n \in \mathbb{N}.$$

Hence

$$\sum_{k=1}^{\infty} z_k^2 < \infty.$$

It is proved that  $z \in \ell_2$ . The fact that  $\{x_n\}$  converges for  $n \rightarrow \infty$  to  $z$  in  $\ell_2$  follows from (1).  $\square$

The next problem addresses the question of the power of different metrics on the same space of functions in terms of convergence. We deal with the space  $\mathcal{S}_c$  of piecewise continuous functions with compact support, equipped with the  $L_p$  metrics. We write briefly  $\mathcal{L}_p$  for these metric spaces. In particular, we show that convergence in  $\mathcal{L}_p$  for some positive  $p$  does not always imply convergence in  $\mathcal{L}_q$  for another positive  $q \neq p$ .

**7.C.17.** Let  $0 < p < \infty$ . For each positive integer  $n$ , define the sequence of functions

$$f_n(x) = \begin{cases} n^{1/p} & -1/n \leq x \leq 1/n \\ 0 & \text{otherwise.} \end{cases}$$

Decide for which  $q$  the sequence  $f_n$  converges in  $\mathcal{L}_q$ .

**Solution.** Let  $q$  be any positive real number. Then

$$\int_{-\infty}^{\infty} |f_n(x)|^q dx = \int_{-1/n}^{1/n} n^{q/p} dx = (2/n)n^{q/p} = 2n^{(q/p)-1}.$$

If  $0 < q < p$  and if  $n \rightarrow \infty$ , then

$$\int_{-\infty}^{\infty} |f_n(x)|^q dx \rightarrow 0.$$

So  $\|f_n\|_q \rightarrow 0$ , and the sequence converges to the zero function. Similarly if  $0 < p < q$  and if  $n \rightarrow \infty$ , then

$$\int_{-\infty}^{\infty} |f_n(x)|^q dx \rightarrow \infty.$$

## ISOMETRIES

A mapping  $\varphi : X_1 \rightarrow X_2$  between metric spaces with metrics  $d_1$  and  $d_2$ , respectively, is called an *isometry* if and only if all elements  $x, y \in X$  satisfy  $d_2(\varphi(x), \varphi(y)) = d_1(x, y)$ .

Of course, every isometry is a bijection onto its image (this follows from the property that the distance between distinct elements is non-zero) and the corresponding inverse mapping is an isometry as well.

Now, consider two inclusions of a dense subset,  $\iota_1 : X \rightarrow \tilde{X}_1$  and  $\iota_2 : X \rightarrow \tilde{X}_2$ , into two completions of the space  $X$ , and denote the corresponding metrics by  $d, d_1$ , and  $d_2$ , respectively. The mapping

$$\varphi : \iota_1(X) \xrightarrow{\iota_1^{-1}} X \xrightarrow{\iota_2} \tilde{X}_2$$

is well-defined on the dense subset  $\iota_1(X) \subset \tilde{X}_1$ . Its image is the dense subset  $\iota_2(X) \subset \tilde{X}_2$  and, moreover, this mapping is clearly an isometry. The dual mapping  $\iota_1 \circ \iota_2^{-1}$  works in the same way.

Every isometric mapping maps, of course, Cauchy sequences to Cauchy sequences. At the same time, such Cauchy sequences converge to the same element in the completion if and only if this holds for their images under the isometry  $\varphi$ . Thus if such a mapping  $\varphi$  is defined on a dense subset  $X$  of a metric space  $\tilde{X}_1$ , then it has a unique extension to the whole  $\tilde{X}_1$  with values lying in the closure of the image  $\varphi(X)$ , i. e.  $\tilde{X}_2$ .

By using the previous ideas, there is a unique extension of  $\varphi$  to the mapping  $\tilde{\varphi} : \tilde{X}_1 \rightarrow \tilde{X}_2$  which is both a bijection and an isometry. Thus, the completions  $\tilde{X}_1$  and  $\tilde{X}_2$  are indeed identical in this sense.

Thus it is proved:

**Theorem.** Let  $X$  be a metric space with metric  $d$  which is not complete. Then the completion  $\tilde{X}$  of  $X$  with metric  $\tilde{d}$  is unique up to bijective isometries.

In the following three paragraphs, we introduce three theorems about complete metric spaces. They are highly applicable in both mathematical analysis and verifying convergence of numerical methods.

**7.3.9. Banach's contraction principle.** A mapping  $F : X \rightarrow X$  on a metric space  $X$  with metric  $d$  is called a *contraction mapping* if and only if there is a real constant  $0 \leq C < 1$  such that for all elements  $x, y$  in  $X$ ,

$$d(F(x), F(y)) \leq C d(x, y).$$

**Theorem.** If  $F$  is a contraction mapping on a complete metric space  $X$ , then it has a fixed point, i. e., there is a  $z \in X$  such that  $F(z) = z$ .

**PROOF.** The proof naturally follows the intuitive idea that iterative application of a contraction mapping starting from an initial value  $z_0 \in X$  should "accumulate" to some point.



So  $\|f_n\|_q$  diverges, and in particular,  $f_n$  cannot converge to any limit.

Finally, for  $q = p$  we have  $\int |f_n(x)|^p dx = 2$  for all positive integers  $n$ , and so as  $n \rightarrow \infty$  we get  $\|f_n\|_p \rightarrow 2^{1/p}$ . At the same time, for any  $g \in \mathcal{S}_c$ , if  $g(x) \neq 0$  at some  $x \neq 0$  where  $g$  is continuous, its distance from  $f_n$  cannot converge to zero.

It follows that  $f_n$  converges to 0 in  $\mathcal{L}_q$ ,  $0 < q < p$ , but it does not converge in  $\mathcal{L}_q$  with  $q \geq p$ .  $\square$

The next problem deals with the extremely useful Banach fixed point theorem, showing the necessity of all the requirements in Theorem 7.3.9.

**7.C.18.** Show that the mapping  $f : \langle 0, \infty \rangle \rightarrow \langle 0, \infty \rangle$  given by

$$f(x) = x + e^{-x}$$

satisfies, for all  $x \neq y$ , the condition

$$|f(x) - f(y)| < |x - y|,$$

but it does not have any fixed point, i.e.  $f(x) \neq x$  for any  $x$ . (Thus the condition  $|f(x) - f(y)| \leq C|x - y|$ , with constant  $C < 1$ , in the Banach fixed point Theorem 7.3.9 is essential.)

**Solution.** Clearly the function  $f$  is strictly increasing on the entire domain. Assume  $y < x$ . Then  $e^{-x} < e^{-y}$  and

$$|f(x) - f(y)| = |x - y + e^{-x} - e^{-y}| < |x - y|.$$

Finally,  $f(x) = x$  implies  $e^{-x} = 0$  which is impossible.  $\square$

Dealing with convergence of real numbers, we observe



that the topological concepts of open neighbourhoods and the compactness are most useful. This is of course even more true for metric spaces, where we can work with open balls of radius  $r$  etc. The definitions remain essentially the same.

**7.C.19.** We have already seen the discrete metric space  $X \neq \emptyset$  with the metric  $d : X \times X \rightarrow \mathbb{R}$  defined by the formula

$$d(x, y) := 1, \quad x \neq y, \quad d(x, y) := 0, \quad x = y.$$

- Decide whether  $(X, d)$  is complete.
- Describe all open, closed, and bounded sets in  $(X, d)$ .
- Describe the interior, boundary, limit, and isolated points of an arbitrary set in  $(X, d)$ .
- Describe all compact sets in  $(X, d)$ .

**Solution.** The case (a) was essentially dealt with in 7.C.12. For an arbitrary sequence  $\{x_n\}_{n \in \mathbb{N}}$  to be a Cauchy sequence, it is necessary in this space that there is an index  $n \in \mathbb{N}$  such that  $x_n = x_{n+m}$  for all  $m \in \mathbb{N}$ . Any sequence with this property then necessarily converges to the common value  $x_n = x_{n+1} = \dots$  (we talk about almost stationary sequences). So the metric space  $(X, d)$  is complete.

The case (b). The open 1-neighbourhood of any element contains this element only. Therefore, every singleton set is open. Since the union of any number of open sets is an open set, every set is open in  $(X, d)$ . By complements,

The metric space  $X$ , of course, needs to be complete; otherwise it could happen that the limit point does not exist in it.

Choose an arbitrary  $z_0 \in X$  and consider the sequence  $z_i, i = 0, 1, \dots$

$$z_1 = F(z_0), \quad z_2 = F(z_1), \dots, \quad z_{i+1} = F(z_i), \dots$$

From the assumptions, we have

$$\begin{aligned} d(z_{i+1}, z_i) &= d(F(z_i), F(z_{i-1})) \\ &\leq C d(z_i, z_{i-1}) \leq \dots \leq C^i d(z_1, z_0). \end{aligned}$$

The triangle inequality then implies that for all natural numbers  $j$ ,

$$\begin{aligned} d(z_{i+j}, z_i) &\leq \sum_{k=1}^j d(z_{i+k}, z_{i+k-1}) \\ &\leq \sum_{k=1}^j C^{i+k-1} d(z_1, z_0) = C^i d(z_1, z_0) \sum_{k=1}^j C^{k-1} \\ &\leq C^i d(z_1, z_0) \sum_{k=1}^{\infty} C^{k-1} = \frac{C^i}{1-C} d(z_1, z_0). \end{aligned}$$

Now, since  $0 \leq C < 1$ ,  $\lim_{n \rightarrow \infty} C^n = 0$ , so for every positive (no matter how small)  $\varepsilon$ , the right-hand expression is surely less than  $\varepsilon$  for sufficiently large indices  $i$ , that is,

$$d(z_i, z_{i+j}) \leq \frac{C^i}{1-C} d(z_1, z_0) \leq \varepsilon.$$

However, this ensures that the sequence  $z_i$  is a Cauchy sequence. Since  $X$  is complete, the sequence has a limit  $z$ , and all that remains to be proved is  $F(z) = z$ .

Every contraction mapping is continuous. Therefore,

$$F(z) = F\left(\lim_{n \rightarrow \infty} z_n\right) = \lim_{n \rightarrow \infty} F(z_n) = z.$$

This finishes the proof.  $\square$

The next two theorems extend the intuitive understanding of “density” of closed intervals  $[a, b] \subset \mathbb{R}$ , not allowing for any “holes” there. They are essential for the understanding of compactness of metric spaces. In fact, they are both special cases of more general theorems on topological spaces.



**7.3.10. Cantor intersection theorem.** For any set  $A$  in a metric space  $X$  with metric  $d$ , the real number

$$\text{diam } A = \sup_{x, y \in A} d(x, y)$$

is called the *diameter of the set A*. The set  $A$  is said to be *bounded* if and only if  $\text{diam } A < \infty$ .

**Theorem.** If  $A_1 \supset A_2 \supset \dots \supset A_i \supset \dots$  is a non-increasing sequence of non-empty closed subsets in a complete metric space  $X$  and if  $\text{diam } A_i \rightarrow 0$ , then there is exactly one point  $x \in X$  belonging to the intersection of all the sets  $A_i$ .<sup>2</sup>

<sup>2</sup>Georg Cantor is considered as the founder of the set theory which he introduced and developed in the last quarter of 19th century. At this time,

this also means that every set is closed. The fact that the 2-neighbourhood of any element coincides with the whole space implies that every set is bounded in  $(X, d)$ .

The case (c). Once again, we use the fact that the open 1-neighbourhood of any element contains this element only. It follows that every point of any set is both its interior and its isolated point and that the sets have neither boundary nor limit points.

The case (d). Every finite set in an arbitrary metric space is compact (it defines a compact metric space by restricting the domain of  $d$ ). It follows from the classification of convergent sequences (see (a)), that no infinite sequence can be compact in  $(X, d)$ .  $\square$

**7.C.20.** In the metric space  $\mathcal{S}[-1, 1]$  with metric given by the norm  $\|\cdot\|_\infty$ , consider the sets

$$A = \{f \in \mathcal{S}[-1, 1]; f(0) \in (0, 2)\},$$

$$B = \{f \in \mathcal{S}[-1, 1]; \int_{-1}^1 f(x) \, dx = 0\}.$$

Are these sets open? Are these sets closed?

**Solution.** The interior of a set  $M$  is the set of all interior points of  $M$  and it is usually denoted by  $M^0$ . A set  $M$  is then open if and only if  $M = M^0$ . Similarly, we define the closure of a set  $M$  as the set of all points having zero distance from  $M$ ; it is denoted by  $\overline{M}$ . A set  $M$  is closed if and only if  $M = \overline{M}$ . Since

$$A^0 = A, \quad \overline{A} = \{f \in \mathcal{S}[-1, 1]; f(0) \in [0, 2]\}, \quad B^0 = \emptyset, \quad \overline{B} = B$$

(especially,  $\overline{A}$  contains functions  $f$  for which  $f(0)$  can attain values from the whole closed interval  $[0, 2]$ ), the set  $A$  is open and not closed, and, the other way around, the set  $B$  is closed and not open.  $\square$

One of the most important concepts related to complete metric spaces is given by the principle of nested balls (cf. the theorem 7.3.10). Under some additional conditions, it says that a metric space  $(X, d)$  is complete if and only if every sequence  $\{A_n\}_{n \in \mathbb{N}}$  of nested (i.e.,  $A_{n+1} \subseteq A_n, n \in \mathbb{N}$ ) non-empty closed sets  $A_n$  has non-empty intersection. That is,

$$(1) \quad \bigcap_{n \in \mathbb{N}} A_n \neq \emptyset.$$

**7.C.21.** Verify that the additional condition in the theorem

$$(1) \quad \lim_{n \rightarrow \infty} \sup \{d(x, y); x, y \in A_n\} = 0.$$

cannot be omitted.

**Solution.** That the requirement (1) cannot be omitted is probably contrarily to many readers' expectations. For a counterexample, consider the set  $X = \mathbb{N}$  with metric

$$d(m, n) = 1 + \frac{1}{m+n}, \quad m \neq n, \quad d(m, n) = 0, \quad m = n.$$

It is indeed a metric. The first and second properties are clearly satisfied. To prove the triangle inequality, it suffices to observe that  $d(m, n) \in (1, 4/3]$  if  $m \neq n$ . Hence the only

**PROOF.** Select one point  $x_i$  for each set  $A_i$ . Since  $\text{diam } A_i \rightarrow 0$ , for every positive real number  $\varepsilon$ , we can find an index  $n(\varepsilon)$  such that for all  $A_i$  with indices  $i \geq n(\varepsilon)$ , their diameters are less than  $\varepsilon$ . For sufficiently large indices  $i, j$ ,  $d(x_i, x_j) \leq \varepsilon$ , and thus our sequence is a Cauchy sequence. Therefore, it has a limit point  $x \in X$ .  $x$  must be a limit point of all the sets  $A_i$ , thus it belongs to all of them (since they are all closed). So  $x$  belongs to their intersection. This proves the existence of  $x$ .

Assume there are two points  $x$  and  $y$ , both belonging to the intersection of all the sets  $A_i$ . Then  $d(x, y)$  must be less than the diameter of the sets  $A_i$ . But  $\text{diam } A_i \rightarrow 0$ , so  $d(x, y) = 0$ , hence  $x = y$ . This proves the uniqueness of  $x$ .  $\square$

**7.3.11. Theorem (Baire theorem).** *If  $X$  is a complete metric space, then the intersection of every countable system of open dense sets  $A_i$  is a dense set in the metric space  $X$ .*<sup>3</sup>

**PROOF.** Suppose  $X$  contains a system of dense open sets  $A_i, i = 1, 2, \dots$ . It is required to show that the set  $A = \bigcap_{i=1}^\infty A_i$  has a non-empty intersection with any open set  $U \subset X$ . Proceed inductively, invoking the previous theorem.

Since  $A_1$  is dense, there is a  $z_1 \in A_1 \cap U$ , but since the set  $A_1$  is open, the closure of an  $\varepsilon_1$ -neighbourhood  $U_1$  (for sufficiently small  $\varepsilon_1$ ) of the point  $z_1$  is contained in  $A_1$  as well. Denote the closure of this  $\varepsilon_1$ -ball  $U_1$  by  $B_1$ .

Further, suppose that the points  $z_i$  and their open  $\varepsilon_i$ -neighbourhoods  $U_i$  are already chosen for  $i = 1, \dots, n$ , so that

$$z_k \in B_k = \overline{U}_k, \quad z_k \in \bigcap_{i=1}^k B_i \subset \bigcap_{i=1}^k A_i \cap U.$$

Since the set  $A_{n+1}$  is open and dense in  $X$ , there is a point  $z_{n+1} \in A_{n+1} \cap U_n$ ; however, since  $A_{n+1} \cap U_n$  is open, the point  $z_{n+1}$  belongs to it together with a sufficiently small  $\varepsilon_{n+1}$ -neighbourhood  $U_{n+1}$ .

Then, the closures surely satisfy  $B_{n+1} = \overline{U}_{n+1} \subset \overline{U}_n$ , and so the closed set  $B_{n+1}$  is contained in  $A_{n+1} \cap \overline{U}_n$ . Moreover, we can assume that all  $\varepsilon_n \leq 1/n$ .

If we proceed in this inductive way from the original point  $z_1$  and the set  $B_1$ , we obtain a non-increasing sequence of non-empty closed sets  $B_n$  whose diameter approaches zero. Therefore, there is a point  $z$  common to all of these sets. That is,

$$z \in \bigcap_{i=1}^\infty \overline{U}_i = \bigcap_{i=1}^\infty B_i \subset \bigcap_{i=1}^\infty A_i \cap U,$$

which is the statement to be proved.  $\square$

the new abstract approach to fundamentals of Mathematics caused fierce objections. It also led to the severe internal crises of Mathematics in the beginning of the 20th century. This part of the history of Mathematics is fascinating.

<sup>3</sup>This theorem is a part of considerations by René-Louis Baire in his 1899 doctoral thesis. More generally, a topological space satisfying the property as in the theorem is called a *Baire space* and the theorem simply says that every complete metric space is a Baire space.

Cauchy sequences are those which are constant from some index on. These sequences are constant except for finitely many terms, sometimes called almost stationary sequences. Thus, every Cauchy sequence is convergent, so the metric space is complete. Define

$$A_n := \left\{ m \in \mathbb{N}; d(m, n) \leq 1 + \frac{1}{2n} \right\}, \quad n \in \mathbb{N}.$$

As the inequality in their definition is not strict, it is guaranteed that they are closed sets. Since  $A_n = \{n, n + 1, \dots\}$ , it follows that  $\{A_n\}$  are nested, but with empty intersection (contrary to (1)). If the requirement (1) is omitted, then the metric space is not complete, contradicting the data. Of course, in this case the condition (1) is not met, as

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \{d(x, y); x, y \in A_n\} \\ &= \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{2n + 1} \right) = 1 \neq 0. \end{aligned}$$

□

**7.C.22.** Determine whether the set (known as the Hilbert cube)

$$A = \left\{ \{x_n\}_{n \in \mathbb{N}} \in \ell_2; |x_n| \leq \frac{1}{n}, n \in \mathbb{N} \right\}$$

is compact in  $\ell_2$ . Then determine the compactness of the set

$$B = \left\{ \{x_n\}_{n \in \mathbb{N}} \in \ell_\infty; |x_n| < \frac{1}{n}, n \in \mathbb{N} \right\}$$

in the space  $\ell_\infty$ .

**Solution.** The space  $\ell_2$  is complete (see 7.C.16). Every closed subset of a complete metric space defines a complete metric space. The set  $A$  is evidently closed in  $\ell_2$ , so it suffices to show that it is totally bounded, and from the theorem 7.3.13(3) it is compact. To do that, construct an  $\varepsilon$ -net of  $A$  for any given  $\varepsilon > 0$ : Begin with the well-known series

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

(see (1)).

For every  $\varepsilon > 0$ , there is an  $n(\varepsilon) \in \mathbb{N}$  satisfying

$$\sqrt{\sum_{k=n(\varepsilon)+1}^{\infty} \frac{1}{k^2}} < \frac{\varepsilon}{2}.$$

From each of the intervals  $[-1/n, 1/n]$  for  $n \in \{1, \dots, n(\varepsilon)\}$ , choose finitely many points  $x_1^n, \dots, x_{m(n)}^n$  so that for any  $x \in [-1/n, 1/n]$  that

$$\min_{j \in \{1, \dots, m(n)\}} |x - x_j^n| < \frac{\varepsilon}{\sqrt{5^n}}.$$

Consider such sequences  $\{y_n\}_{n \in \mathbb{N}}$  from  $\ell_2$  whose terms with indices  $n > n(\varepsilon)$  are zero, and at the same time,

$$y_1 \in \left\{ x_1^1, \dots, x_{m(1)}^1 \right\}, \dots, y_{n(\varepsilon)} \in \left\{ x_1^{n(\varepsilon)}, \dots, x_{m(n(\varepsilon))}^{n(\varepsilon)} \right\}.$$

There are only finitely many such sequences and they create the desired  $\varepsilon$ -net for  $A$ : let  $x_n \in \ell_2$  is arbitrary. According to our choice of the sequences  $y_n$ , there is  $y_n$  such that



**7.3.12. Bounded and compact sets.** The following concepts facilitated our discussions when dealing with the real and complex numbers. They can be reformulated for general metric spaces with almost no change:

An *interior point* of a subset  $A$  in a metric space is such an element of  $A$  which belongs to it together with some of its  $\varepsilon$ -neighbourhoods.

A *boundary point* of a set  $A$  is an element  $x \in X$  such that each its neighbourhood has a non-empty intersection with both  $A$  and the complement  $X \setminus A$ . A boundary point may or may not belong to the set  $A$  itself.

A *limit point* of a set  $A$  is an element  $x$  equal to the limit of a sequence  $x_i \in A$ , such that  $x_i \neq x$  for all  $i$ . Clearly a limit point may or may not belong to the set  $A$ .

An *isolated point* of a set  $A$  is an element  $a \in A$  such that one of its  $\varepsilon$ -neighbourhoods in  $X$  has the singleton intersection  $\{a\}$  with  $A$ .

An *open cover* of a set  $A$  is a system of open sets  $U_i \subset X$ ,  $i \in I$ , such that their union contains  $A$ .

COMPACT SETS

A metric space  $X$  is called *compact* if every sequence of points  $x_i \in X$  has a subsequence converging to some point  $x \in X$ .

Any subset  $A \subset X$  in a metric space is called *compact* if it is compact as the metric space with the restricted metric.

Clearly, the compact subsets in discrete metric spaces  $X$  are exactly the finite subsets of  $X$ .

In the case of the real numbers  $\mathbb{R}$  and complex numbers  $\mathbb{C}$ , our definition reveals the compact subsets discussed there and we would also like to come to useful properties as we did for real and complex numbers in the paragraphs 5.2.7–5.2.8.

Just as an appetizer, notice that the continuous functions between general metric spaces behave as the real functions:

**Theorem.** Let  $f : X \rightarrow Y$  be a continuous mapping between metric spaces. Then the images of compact sets are compact.

**PROOF.** Recall that any convergent sequence of points  $x_i \rightarrow x$  in  $X$  is mapped onto the convergent sequence  $f(x_i) \rightarrow f(x)$  in  $Y$ . Thus, the statement follows immediately from our definition of the compactness via convergent subsequences. □

In particular we obtain the most useful consequence on the minima and maxima of continuous real valued functions on compact subsets:

**Corollary.** Let  $f : X \rightarrow \mathbb{R}$  be a real function defined on a compact metric space. Then there are the points  $x_0$  and  $y_0$  in  $X$  such that

$$f(x_0) = \max_{x \in X} \{f(x)\}, \quad f(y_0) = \min_{x \in X} \{f(x)\}.$$

**PROOF.** The image  $f(X)$  must be a compact subset in  $\mathbb{R}$ , thus it must achieve both maximum and minimum (which are

$$\begin{aligned}
 d(x_n, y_n) &= \sqrt{\sum_{k=1}^{\infty} (x_k - y_k)^2} \\
 &\leq \sqrt{\sum_{k=1}^{n(\varepsilon)} (x_k - y_k)^2} + \sqrt{\sum_{k=n(\varepsilon)+1}^{\infty} x_k^2} \\
 &\leq \sqrt{\frac{\varepsilon^2}{5} + \frac{\varepsilon^2}{5^2} + \cdots + \frac{\varepsilon^2}{5^{n(\varepsilon)}}} + \frac{\varepsilon}{2} \\
 &< \varepsilon \cdot \sqrt{\frac{1}{1 - \frac{1}{5}} - 1} + \frac{\varepsilon}{2} = \varepsilon.
 \end{aligned}$$

Since  $\varepsilon > 0$  is arbitrary, the set  $A$  is totally bounded, which implies compactness.

The closure of the set  $B$  is

$$\bar{B} = \left\{ \{x_n\}_{n \in \mathbb{N}} \in l_{\infty}; |x_n| \leq \frac{1}{n}, n \in \mathbb{N} \right\}.$$

Hence  $B$  is not closed, and so it is not compact. The set  $\bar{B}$  is compact. The proof of this fact is much simpler than for the set  $A$ , thus we leave it as an exercise for the reader.  $\square$

**7.C.23.** Prove that on each metric space  $X$ , the given metric  $d$  is a continuous function  $X \times X \rightarrow \mathbb{R}$ .  $\circ$

**7.C.24.** Show that if  $F$  is a continuous mapping on a compact metric space  $X$ , then the inequality

$$d(F(x), F(y)) < d(x, y),$$

for all  $x \neq y$ , implies the existence of a fixed point.

**Solution.** The infimum  $\alpha$  of the values of the continuous function  $d(x, F(x))$  must be achieved in a point  $x_0 \in X$  (see 7.3.12 for the concepts and main results and use the previous result 7.C.23). Since distances are non-negative,  $\alpha \geq 0$ . If  $\alpha \neq 0$ , then

$$d(F(x_0), F(F(x_0))) < d(x_0, F(x_0)) = \alpha$$

which is a contradiction.  $\square$

the supremum and the infimum of the bounded and closed image).  $\square$

The concept of boundedness is a little more complicated in the case of general metric spaces. For any point  $x$  and subset  $B \subset X$  in a metric space  $X$  with metric  $d$ , we define their *distance*<sup>4</sup>

$$\text{dist}(x, B) = \inf_{y \in B} \{d(x, y)\}.$$

We say that a metric space  $X$  is *totally bounded* if for every positive real number  $\varepsilon$ , there is a finite set  $A$  such that

$$\text{dist}(x, A) < \varepsilon$$

for all points  $x \in X$ . We call such an  $A$  an  $\varepsilon$ -net of  $X$ .

Note that a metric space is *bounded* if  $X$  has a finite diameter. We can immediately see that a totally bounded space is always bounded. Indeed, the diameter of a finite set is finite, and if  $A$  is the set corresponding to  $\varepsilon$  from the definition of total boundedness, then the distance  $d(x, y)$  of two points can always be bounded by the sum of  $\text{dist}(x, A)$ ,  $\text{dist}(y, A)$ , and  $\text{diam } A$ , which is a finite number.

In the case of a metric on a subset of a finite-dimensional Euclidean space, these concepts coincide since the boundedness of a set guarantees the boundedness of all coordinates in a fixed orthonormal basis, and this implies total boundedness. (Verify this in detail by yourselves!)

The next theorem provides the promised very useful alternative characterisations of compactness:

**7.3.13. Theorem.** *The following statements about a metric space  $X$  are equivalent:*

- (1)  $X$  is compact;
- (2) every open covering of  $X$ ,  $X = \cup_{i \in I} U_i$ , contains a finite covering  $X = \cup_{k=1}^n U_{j_k}$ , where all  $j_k \in I$ ;
- (3)  $X$  is complete and totally bounded.

**PROOF.** We show consecutively the implications (1)  $\implies$  (3)  $\implies$  (2)  $\implies$  (1).

Thus, start assuming  $X$  is compact. Then for each Cauchy sequence of points  $x_i$ , there is a sub-sequence  $x_{i_n}$  converging to a point  $x \in X$ . We just have to verify that the initial sequence also converges to the same limit,  $x_i \rightarrow x$ . This is easy and we leave it to the reader. So  $X$  is complete.

Suppose  $X$  is not totally bounded. Then there is  $\varepsilon > 0$  such that no finite  $\varepsilon$ -net exists in  $X$ . Then there is a sequence of points  $x_i$  such that  $d(x_i, x_j) \geq \varepsilon$  for all  $i \neq j$ . (Verify this almost obvious claim – look at the definition of  $\varepsilon$ -nets!) Then this is a sequence of points, where no sub-sequence can be a Cauchy sequence, so  $X$  is not compact. This contradicts (1), so we conclude that  $X$  is totally bounded.

<sup>4</sup>Notice, that the distance between two subsets  $A, B \subset X$  should express how “different” they are. Thus we define the (Hausdorff) distance as follows  $\text{dist}(A, B) = \max\{\sup\{\text{dist}(x, B), x \in A\}, \sup\{\text{dist}(y, A), y \in B\}\}$ . This difference is finite for bounded sets and it is easy to see that it vanishes if and only if the closures of  $A$  and  $B$  coincide. This means, the distances of the point  $x$  and one-point set  $\{x\}$  are very much different,  $\text{dist}(x, A) \neq \text{dist}(\{x\}, A)$ , in general.



The next implication, namely (3)  $\implies$  (2) is more demanding. So assume  $X$  is complete and totally bounded, but  $X$  does not satisfy (2).

Then there is an open covering  $U_\alpha$ ,  $\alpha \in I$ , of  $X$ , which does not contain any finite covering. Choose a sequence of positive real numbers  $\varepsilon_k \rightarrow 0$  and consider the finite  $\varepsilon_k$ -nets from the definition of total boundedness. Further, for each  $k$ , consider the system  $\mathcal{A}_k$  of closed balls with centres in the points of the  $\varepsilon_k$ -net and diameters  $2\varepsilon_k$ . Clearly each such system  $\mathcal{A}_k$  covers the entire space  $X$ . Altogether, there must be at least one closed ball  $C$  in the system  $\mathcal{A}_1$  which is not covered by a finite number of the sets  $U_\alpha$ . Call it  $C_1$  and notice that  $\text{diam } C_1 = 2\varepsilon_1$ .

Next, consider the sets  $C_1 \cap C$ , with balls  $C \in \mathcal{A}_2$  which cover the entire set  $C_1$ . Again, at least one of them cannot be covered by a finite number of  $U_\alpha$ , we call it  $C_2$ . This way, we inductively construct a sequence of sets  $C_k$  satisfying  $C_{k+1} \subset C_k$ ,  $\text{diam } C_k \leq 2\varepsilon_k$ ,  $\varepsilon_k \rightarrow 0$ , and none of them can be covered by a finite number of the open sets  $U_\alpha$ .

Finally we choose one point  $x_k \in C_k$  in each of these sets. By construction, this must be a Cauchy-sequence. Consequently, this sequence of points has a limit  $x$  since  $X$  is complete. Thus there is  $U_{\alpha_0}$  containing  $x$  and containing also some  $\delta$ -neighbourhood  $B_\delta(x)$ . But now, if  $\text{diam } C_k \leq 2\varepsilon_k < \delta$ , then  $C_k \subset B_\delta(x) \subset U_{\alpha_0}$ , which is a contradiction.

The remaining step is to show the implication (2)  $\implies$  (1). Assume (2) and considering any sequence of points  $x_i \in X$ , we set  $C_n = \{x_k; k \geq n\}$ . The intersection of these sets must be non-empty by the following general lemma:

**Lemma.** *Let  $X$  be a metric space such that property (2) in the Theorem holds. Consider a system of closed sets  $D_\alpha$ ,  $\alpha \in I$ , such that each its finite subsystem  $D_{\alpha_1}, \dots, D_{\alpha_k}$  has non-empty intersection. Then also*

$$\bigcap_{\alpha \in I} D_\alpha \neq \emptyset.$$

This simple lemma is proved by contradiction, again. If the latter intersection is empty, then

$$X = X \setminus (\bigcap_{\alpha \in I} D_\alpha) = \bigcup_{\alpha \in I} (X \setminus D_\alpha) = \bigcup_{\alpha \in I} V_\alpha,$$

where  $V_\alpha = X \setminus D_\alpha$  are open sets. Thus, there must be a finite number of them,  $\{V_{\alpha_1}, \dots, V_{\alpha_n}\}$ , covering  $X$  too. Thus, we obtain

$$X = \bigcup_{i=1}^n V_{\alpha_i} = \bigcup_{i=1}^n (X \setminus D_{\alpha_i}) = X \setminus (\bigcap_{i=1}^n D_{\alpha_i}).$$

This is a contradiction with our assumptions on  $D_\alpha$  and the lemma is proved.

Now, let  $x \in \bigcap_{n=1}^\infty C_n$ . By construction, there is a subsequence  $x_{n_k}$  in our sequence of points  $x_n \in X$ , so that  $d(x_{n_k}, x) < 1/k$ . This is a converging subsequence, and so the proof is complete.  $\square$

As an immediate corollary of the latter theorem, each closed subset in a compact metric space is again compact. For subsets of a totally bounded set are totally bounded, and closed subsets of a complete metric space are also complete.

Another consequence is an alternative proof that a subset  $K \subset \mathbb{R}^n$  is compact, if and only if it is closed and bounded.

Notice also that while the conditions (1) and (3) are given in terms of the metric, the equivalent condition (2) is purely topological.

**7.3.14. Continuous functions.** We revisit the questions related to continuity of mappings between metric spaces. In fact, many ideas understood for the functions of one real variable generalize naturally.



In particular, as we already showed, every continuous function  $f : X \rightarrow \mathbb{R}$  on a compact set  $X$  is bounded and achieves its maximum and minimum. Here is another argument for this using the purely topological concept of compactness:

Consider the open intervals  $U_n = (n - 1, n + 1) \subset \mathbb{R}$ ,  $n \in \mathbb{Z}$  covering  $\mathbb{R}$ . Then their preimages  $f^{-1}(U_i)$  cover  $X$ , so that there is a finite number of them, covering  $X$  as well. Thus  $f$  is bounded and the supremum and infimum of its values exist. Consider sequences  $f(x_n)$  and  $f(y_n)$  converging to the supremum and infimum, respectively. Then there must be convergent subsequences of the points  $x_n$  and  $y_n$  in  $X$  and their limits  $x$  and  $y$  are in  $X$  too. But then  $f(x)$  and  $f(y)$  are the supremum and infimum of the values of  $f$  since  $f$  is continuous and thus respects convergence.

We should also enjoy to see the differences between the “purely topological” concepts, as the continuity (possibly defined merely by means of open sets), and the next stronger concepts, which are “metric” properties.

UNIFORMLY CONTINUOUS MAPPINGS

A mapping  $f : X \rightarrow Y$  between metric space is called *uniformly continuous*, if for each  $\varepsilon > 0$  there is a  $\delta > 0$ , such that  $d_Y(f(x), f(y)) < \varepsilon$  for all  $x, y \in X$  with  $d_X(x, y) < \delta$ .

Notice the following simple lemma:

**Lemma.** *A mapping  $f : X \rightarrow Y$  between metric spaces is uniformly continuous if and only if for each pair of sequences  $x_k$  and  $y_k$  in  $X$ ,  $d_X(x_k, y_k) \rightarrow 0$  implies  $d_Y(f(x_k), f(y_k)) \rightarrow 0$ .*

**PROOF.** If  $f$  is uniformly continuous, then the the claim on the sequences is obvious.

Now, assume the property of sequences holds true, but  $f$  is not uniformly continuous. Then, employing the negation of the defining property, there is an  $\varepsilon > 0$  such that for all  $\delta = 1/n$ , there are  $x_n, y_n$  with  $d_X(x_n, y_n) < 1/n$  and  $d_Y(f(x_n), f(y_n)) \geq \varepsilon$ . But this violates the assumed condition. □

This observation leads to the following generalization of the behavior of real functions:

**Theorem.** *Each continuous mapping  $f : X \rightarrow Y$  on a compact metric space  $X$  is uniformly continuous.*

PROOF. Assume  $f$  is a continuous function. Consider any two sequences  $x_k$  and  $y_k$  with  $d_X(x_k, y_k) \rightarrow 0$  and assume their images do not satisfy  $d_Y(f(x_k), f(y_k)) \rightarrow 0$ . Then, for some  $\varepsilon > 0$ , there must be a subsequence  $(x_{k_n}, y_{k_n})$  for which  $d_Y(f(x_{k_n}), f(y_{k_n})) > \varepsilon$ .

Since  $X$  is compact, there is a further subsequence of  $x_{k_n}$  converging to some point  $x \in X$ . Without loss of generality, we may assume that the original sequence  $(x_k, y_k)$  has got all these properties (we are aiming at a conflict with our assumption that  $f$  is not uniform continuous). Of course,  $d_X(x, y_k) \leq d_X(x, x_k) + d_X(x_k, y_k) \rightarrow 0$  and so  $\lim_{k \rightarrow \infty} y_k = x$ , too.

Next, notice that the metric  $d_Y : Y \times Y \rightarrow \mathbb{R}$  is always a continuous function. Indeed, remind the definition of the product metric on  $X \times X$  by the maximum of the distances of the components, see 7.C.1. With this definition, the continuity of  $d$  is obvious from the estimate

$$|d(x, y) - d(x', y')| \leq d(x, x') + d(y, y'),$$

cf. 7.3.8(1).

Now, the continuity of  $f$  ensures

$$\lim_{k \rightarrow \infty} d_Y(f(x_k), f(y_k)) = d_Y(f(x), f(x)) = 0.$$

But this violates our assumption on the sequences and the proof is complete.  $\square$

A very useful variation on the theme of continuity is the following definition.

#### LIPSCHITZ CONTINUITY

A function  $f : X \rightarrow Y$  between metric spaces is called *Lipchitz continuous* if there is a constant  $C > 0$  such that for all points  $x, y \in X$

$$d_Y(f(x), f(y)) \leq C d_X(x, y).$$

Every Lipschitz continuous function is uniformly continuous.

**7.3.15. Arzelà-Ascoli Theorem.** To conclude, we consider some spaces of functions. These provide examples of how much they may differ from the usual Euclidean spaces. First, we introduce some terminology. Basically we want to deal with functions, which are all uniformly continuous in the very same way:

#### EQUICONTINUOUS FUNCTIONS

Consider a space  $M$  of mappings  $f : X \rightarrow Y$  between metric spaces. We say that the functions in  $M$  are *equicontinuous*, if for each  $\varepsilon > 0$  there is a  $\delta > 0$ , such that  $d_Y(f(x), f(y)) < \varepsilon$  for all  $x, y \in X$  with  $d_X(x, y) < \delta$ , for all functions  $f \in M$ .

Consider the metric space  $\mathcal{C}(X)$  of all continuous (real or complex) functions on a compact metric space  $X$ , with its  $\|\cdot\|_\infty$  norm. This means that the distance between two functions  $f, g$  is the maximum of the distance between their values  $f(x)$  and  $g(x)$  for  $x \in X$ .



We say that a set  $M \subset \mathcal{C}(X)$  of real functions is *uniformly bounded*, if there is a constant  $K \in \mathbb{R}$  such that  $|f(x)| \leq K$  for all functions  $f \in M$  and points  $x \in X$ . Of course, bounded sets  $M$  of functions in  $\mathcal{C}(X)$  are always uniformly bounded, by the definition of the norm.

**Theorem.** Consider a compact metric space  $X$ . A set  $M \subset \mathcal{C}(X)$  in the space of continuous functions with the supremum norm  $\|\cdot\|_\infty$  is compact if and only if it is bounded, closed, and equicontinuous.<sup>5</sup>



**PROOF.** Suppose  $M$  is compact. Then  $M$  is totally bounded (and thus also uniformly bounded as noticed above). Since every compact subset is closed, it remains to verify the equicontinuity.

Given  $\varepsilon > 0$ , consider the corresponding  $\varepsilon$ -net  $(f_1, f_2, \dots, f_k) \subset M$  from the definition of the total boundedness of  $M$ . Recall that all the functions  $f_i$  are uniformly continuous (as continuous functions on a compact set). Thus there is a  $\delta_i$  for each  $f_i$ , such that  $d_X(x, y) < \delta_i$  implies  $|f_i(x) - f_i(y)| < \varepsilon$ .

Of course, we take  $\delta$  to be the minimum of the finite many  $\delta_i$ ,  $i = 1, \dots, k$ . Then the same equality holds for all  $f_i$  in the  $\varepsilon$ -net. But now, considering an arbitrary function  $f \in M$ , there is a function  $f_j$  in our  $\varepsilon$ -net with  $\|f - f_j\|$  and so  $d_X(x, y) < \delta$  implies that  $|f(x) - f(y)|$  is at most

$$|f(x) - f_j(x)| + |f_j(x) - f_j(y)| + |f_j(y) - f(y)| \leq 3\varepsilon,$$

and the equicontinuity has been proved.

Conversely, suppose that  $M$  is a bounded, closed, and equicontinuous subset of  $\mathcal{C}(X)$ , with  $X$  as a compact metric space.



First we show that  $M$  is complete. Obviously, a Cauchy sequence in the  $\|\cdot\|_\infty$  norm is also pointwise Cauchy sequence of real or complex numbers. Thus the sequence of functions converges pointwise. Next, the pointwise limit must be bounded and the equicontinuity property implies that this limit will be continuous as well.

Thus, we need to find a Cauchy (sub)sequence within any sequence of functions  $f_n \in M$ .

The compact space  $X$  itself is totally bounded and therefore it contains a countably dense set  $A \subset X$  (we may take the points in all  $1/k$ -nets for  $k \in \mathbb{N}$ ). Write  $A = \{a_1, a_2, \dots\}$  as a sequence.

Choose a subsequence of functions  $f_{1j}$ ,  $j = 1, 2, \dots$  within the functions  $f_n$ , so that the sequence of values  $f_{1j}(a_1)$  converges. (This is possible, since the set  $M$  is bounded in the  $\|\cdot\|_\infty$  norm). Similarly, the subsequence  $f_{2j}$  can be chosen from  $f_{1k}$ , so that  $f_{2j}(a_2)$  converges. In general, the  $m$ -th subsequence is chosen from  $f_{(m-1)k}$  and have the values  $f_{mj}(a_m)$  converging (and by our construction, it converges in all  $a_i$ ,  $i < m$  too).

<sup>5</sup>A weaker version providing a sufficient condition was first published by Ascoli in 1883, the complete exposition was given by Arzelà in 1895. Again, there are much more general versions of this theorem in the realm of topological spaces.

As a result, we can choose the sequence of function  $g_k = f_{kk}$  for all positive integers  $k$  with the hope that this is a Cauchy sequence. This is where the equicontinuity helps.

Start with any  $\varepsilon > 0$  and find  $\delta_\varepsilon > 0$ , such that  $|f(x) - f(y)| < \varepsilon$  whenever the arguments  $x$  and  $y$  are closer than  $\delta_\varepsilon$ . Let  $A_\varepsilon \subset A$  be subset forming a  $\delta_\varepsilon$ -net. This is a finite set and so there must be an  $n \in \mathbb{N}$  such that for all  $i, j \geq n$  and all  $a \in A_\varepsilon$ , we know  $|g_i(a) - g_j(a)| < \varepsilon$ . But then, for every  $x \in X$ , there is some  $a \in A_\varepsilon$  with  $d_X(x, a) < \delta_\varepsilon$  and so  $|g_i(x) - g_j(x)|$  can be at most

$$|g_i(x) - g_i(a)| + |g_i(a) - g_j(a)| + |g_j(a) - g_j(x)| \leq 3\varepsilon.$$

Thus, the sequence  $g_k$  is a Cauchy sequence in  $C(X)$ , and so  $M$  is compact.  $\square$

**D. Additional exercises to the whole chapter**

Let us begin this section with problems similar to those described in the introduction of Chapter 7. (cf. 7.A.1, 7.A.2).

**7.D.1.** Let  $\mathcal{S}^0[0, 1]$  be the space of  $\mathbb{F}$ -valued piecewise continuous functions defined on  $[0, 1]$ , where  $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ , endowed with the  $L_2$ -inner product



$$\langle f, g \rangle := \int_0^1 f(x)\overline{g(x)} dx.$$

With respect to the induced norm  $\|\cdot\|_2$  compute:

- (a)  $\|f\|_2$ , where  $f(x) = x + b$ , with  $b$  being some constant;
- (b) the distance  $\|f - g\|_2$  between  $f(x) = \cos(2\pi x)$  and  $g(x) = \sin(2\pi x)$ .

**Solution.** (a) By definition,  $\|f\|_2 = \sqrt{\langle f, f \rangle} = \sqrt{\int_0^1 (x + b)^2 dx}$ , and we easily compute that

$$\int_0^1 (x + b)^2 dx = \left[ \frac{x^3}{3} \right]_0^1 + 2 \left[ \frac{bx^2}{2} \right]_0^1 + [b^2x]_0^1 = \frac{1}{3} + b + b^2.$$

Thus  $\|f\|_2 = \sqrt{\frac{1}{3} + b + b^2}$ .

For (b) and the distance between the given  $f, g$ , we compute

$$\begin{aligned} \langle f - g, f - g \rangle &= \int_0^1 (\cos(2\pi x) - \sin(2\pi x))^2 dx = \int_0^1 (1 - 2\cos(2\pi x)\sin(2\pi x)) dx = \int_0^1 (1 - \sin(4\pi x)) dx \\ &= \left[ x - \frac{\cos(4\pi x)}{4\pi} \right]_0^1 = 1, \end{aligned}$$

where the third equality relies on the identity  $\sin(2t) = 2\sin(t)\cos(t)$ . Thus  $\|f - g\|_2 = \sqrt{1} = 1$ . □

**7.D.2.** Compute the  $L_2$ -norm of the function  $f(x) = e^{ix}$ , defined on the interval  $[-\pi, \pi]$ .

**Solution.** We can express  $f$  as  $f(x) = e^{ix} = \cos(x) + i\sin(x)$ , for all  $x \in [-\pi, \pi]$ . Then, with respect to  $L_2$ -product we see that

$$\langle f, f \rangle = \int_{-\pi}^{\pi} (\cos(x) + i\sin(x))(\cos(x) - i\sin(x)) dx = \int_{-\pi}^{\pi} (\cos^2(x) + \sin^2(x)) dx = \int_{-\pi}^{\pi} dx = 2\pi.$$

Thus  $\|f\|_2 = \sqrt{2\pi}$ . □

**7.D.3.** Consider the vector space  $\mathbb{R}_3[x]$  of polynomials of one variable of degree at most 3 with real coefficients, endowed with the  $L_2$ -inner product (i.e., the same inner product  $\langle \cdot, \cdot \rangle$  as the one in Problem 7.D.1). Let  $W$  be the subspace of  $\mathbb{R}_3[x]$  having as basis the set  $B = \{x, x^2\}$ . Find an orthonormal basis of  $W$  with respect to  $\langle \cdot, \cdot \rangle$ .

**Solution.** Set  $E_1 = x$  and  $E_2 = x^2$ . The Gram-Schmidt procedure states that an orthogonal basis of  $W$  is given by

$$\left\{ w_1 = E_1, \quad w_2 = E_2 - \frac{\langle E_2, w_1 \rangle}{\|w_1\|_2^2} w_1 \right\}.$$

We compute

$$\langle E_2, w_1 \rangle = \int_0^1 x^3 dx = \frac{x^4}{4} \Big|_0^1 = \frac{1}{4}, \quad \text{and} \quad \langle w_1, w_1 \rangle = \int_0^1 x^2 dx = \frac{x^3}{3} \Big|_0^1 = \frac{1}{3}.$$

Thus  $w_2 = E_2 - \frac{3}{4}E_1 = x^2 - \frac{3}{4}x$ . As a verification, observe that

$$\langle w_1, w_2 \rangle = \langle x, x^2 - \frac{3}{4}x \rangle = \frac{1}{4} - \frac{3}{4} \cdot \frac{1}{3} = 0.$$

Consequently, an orthonormal basis of  $W$  is given by the “vectors”  $\{\hat{w}_1 = \frac{w_1}{\|w_1\|_2}, \hat{w}_2 = \frac{w_2}{\|w_2\|_2}\}$ , where  $\|w_1\|_2 = \frac{1}{\sqrt{3}}$  and

$$\|w_2\|_2 = \sqrt{\int_0^1 \left(x^2 - \frac{3}{4}x\right)^2 dx} = \frac{1}{\sqrt{80}},$$

respectively. □

**7.D.4.** Consider the inner vector space  $(\mathbb{R}_3[x], \langle \cdot, \cdot \rangle)$  of the previous task 7.D.3, and let  $W$  be the subspace of  $\mathbb{R}_3[x]$  spanned by the basis  $\{1, x^2\}$ . Find a basis of the orthogonal complement  $W^\perp$  of  $W$  with respect to  $\langle \cdot, \cdot \rangle$ .

**Solution.** By assumption  $W$  is spanned by the set  $\{1, x^2\}$ . Thus, a vector  $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0 \in W^\perp$  in  $W^\perp$  must satisfy

$$\langle f(x), 1 \rangle = 0, \quad \text{and} \quad \langle f(x), x^2 \rangle = 0.$$

This gives the following system of equations

$$\left\{ \frac{a_3}{4} + \frac{a_2}{3} + \frac{a_1}{2} + a_0 = 0, \quad \frac{a_3}{6} + \frac{a_2}{5} + \frac{a_1}{4} + \frac{a_0}{3} = 0 \right\}.$$

We can solve this in Sage via the cell

```
var("a3", "a2", "a1", "a0")
solve([(1/4)*a3+(1/3)*a2+(1/2)*a1+a0==0, (1/6)*a3+(1/5)*a2+(1/4)*a1+(1/3)*a0==0], a3, a2, a1, a0)
```

Sage returns the solution

```
[a3 == 16*r1 + 3*r2, a2 == -15*r1 - 15/4*r2, a1 == r2, a0 == r1]
```

which means that  $a_3 = 16a + 3b$  and  $a_2 = -15(a + \frac{b}{4})$  with  $a_1 = b \in \mathbb{R}$  and  $a_0 = a \in \mathbb{R}$  being the free variables. Thus

$$f(x) = b(3x^3 - \frac{15}{4}x^2 + x) + a(16x^3 - 15x^2 + 1),$$

and this implies that  $W^\perp = \text{span}\{f_1(x), f_2(x)\}$ , where  $f_1(x) = 3x^3 - \frac{15}{4}x^2 + x$  and  $f_2(x) = 16x^3 - 15x^2 + 1$ , respectively. It remains to prove that  $f_1(x), f_2(x)$  are linearly independent which one can easily check (for instance one may compute the Wronskian as in the Problem 6.D.16). Thus  $\{f_1, f_2\}$  is a basis of  $W^\perp$ .  $\square$

**7.D.5.** Given the vector subspace  $\langle \sin(x), x \rangle$  of the space of continuous real-valued functions defined on the interval  $[0, \pi]$ , complete the function  $x$  to an orthogonal basis of the subspace and determine the orthogonal projection of the function  $\frac{1}{2} \sin(x)$  onto it.  $\circ$

**7.D.6.** Given the vector subspace  $\langle \cos(x), x \rangle$  of the space of continuous real-valued functions defined on the interval  $[0, \pi]$ , complete the function  $\cos(x)$  to an orthogonal basis of the subspace and determine the orthogonal projection of the function  $\sin(x)$  onto it.  $\circ$

**7.D.7.** Show that the system of functions  $1, \sin(x), \cos(x), \dots, \sin(nx), \cos(nx), \dots$  is orthogonal with respect to the  $L_2$  inner product on the interval  $I = [-\pi, \pi]$ .  $\circ$

*Next we will focus on tasks about Fourier series of periodic functions, and their convergence.*

**7.D.8.** Let  $f : (-\pi, \pi) \rightarrow \mathbb{R}$  be the piecewise function defined by

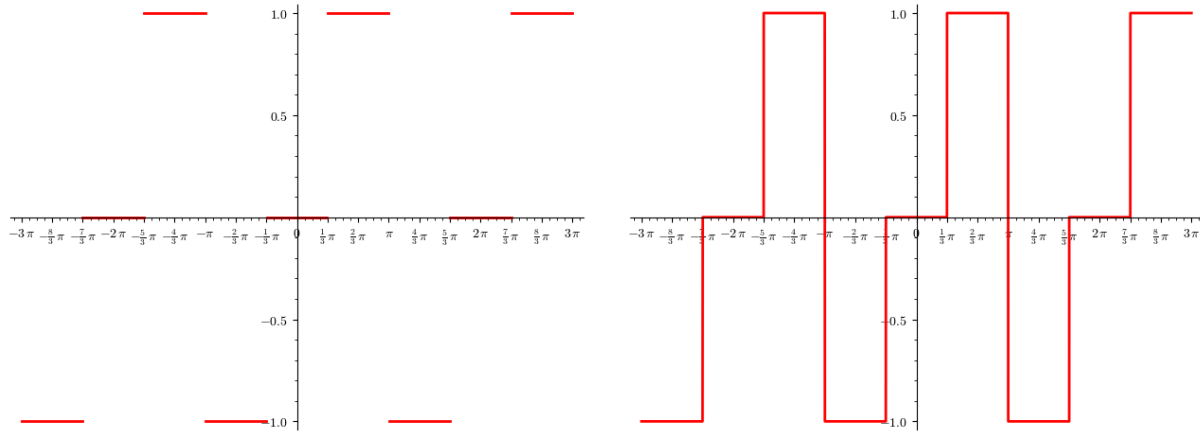
$$f(x) = \begin{cases} -1, & \text{if } -\pi < x < -\pi/3, \\ 0, & \text{if } -\pi/3 \leq x \leq \pi/3, \\ 1, & \text{if } \pi/3 < x < \pi. \end{cases}$$

- (1) Find the  $2\pi$ -periodic extension of  $f$  and sketch its graph via Sage for  $x \in (-3\pi, 3\pi)$ ;
- (2) Prove that the Fourier series  $F$  of the  $2\pi$ -periodic extension of  $f$  is a sine series. Next determine explicitly  $F$ ;
- (3) Use Sage to verify your computations in (2);
- (4) Present the 5th partial sum of the approximation. Next confirm your result via Sage and plot  $F_5, F_{10}, F_{20}$  and  $F_{100}$  together with  $f$ , for  $x \in (-\pi, \pi)$ .
- (5) Use Sage to specify the 5th, 10th, and 100th coefficient of the Fourier series  $F$  corresponding to  $f$ .

**Solution.** (1) The  $2\pi$ -periodic extension of  $f$  is defined by  $\hat{f}(x) = f(x)$  for all  $-\pi < x < \pi$  with  $\hat{f}(x) = \hat{f}(x + 2\pi)$  for all  $x \in \mathbb{R}$ . In Sage to plot this periodic extension for  $x \in (-3\pi, 3\pi)$  we can use appropriately the `piecewise` function. In particular, the block

```
f1(x)=-1; f2(x)=0; f3(x)=1
f=piecewise([( (-3*pi, -7*pi/3), f1), [(-7*pi/3, -5*pi/3), f2], [(-5*pi/3, -pi), f3], [(-pi, -pi/3), f1],
[(-pi/3, pi/3), f2], [(pi/3, pi), f3], [(pi, 5*pi/3), f1], [(5*pi/3, 7*pi/3), f2], [(7*pi/3, 3*pi), f3]])
a = f.plot(x, -3*pi, 3*pi, thickness=2, color="red",
exclude=[-7*pi/3, -5*pi/3, -pi, -pi/3, pi/3, pi, 5*pi/3, 7*pi/3])
a.show(ticks=pi/3, tick_formatter=pi)
```

produces the following figure:



Here one may avoid to use the `exclude` option (to exclude the jump discontinuities) and in this case Sage will produce the figure given in the r.h.s, where the vertical lines should not be considered as part of the required graph.

(2) It is easy to see that  $f(x)$  (and hence  $\hat{f}$ ) is an odd function, thus  $a_n = 0$  for all  $n \in \mathbb{N}$ . Consequently the Fourier series of  $f$  is a sine series. As for  $b_n$ , first notice that the function  $g(x) := f(x) \sin(nx)$  satisfies

$$g(x) = \begin{cases} -\sin(nx), & \text{if } -\pi < x < -\pi/3, \\ 0, & \text{if } -\pi/3 \leq x \leq \pi/3, \\ \sin(nx) & \text{if } \pi/3 < x < \pi, \end{cases} \quad \text{and} \quad g(-x) = \begin{cases} \sin(nx), & \text{if } \pi/3 < x < \pi, \\ 0, & \text{if } -\pi/3 \leq x \leq \pi/3, \\ -\sin(nx), & \text{if } -\pi < x < -\pi/3. \end{cases}$$

Thus  $g(-x) = g(x)$  for all  $x \in (-\pi, \pi)$ , that is,  $g$  is an even function. Therefore,

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(nx) dx = \frac{2}{\pi} \left[ \int_0^{\pi/3} f(x) \sin(nx) dx + \int_{\pi/3}^{\pi} f(x) \sin(nx) dx \right] \\ &= \frac{2}{\pi} \int_{\pi/3}^{\pi} 1 \cdot \sin(nx) dx = \frac{2}{\pi} \left[ -\frac{\cos(nx)}{n} \right]_{\pi/3}^{\pi} = \frac{2}{n\pi} \left( -\cos(n\pi) + \cos\left(\frac{n\pi}{3}\right) \right) = \frac{2}{n\pi} \left( -(-1)^n + \cos\left(\frac{n\pi}{3}\right) \right) \\ &= \frac{2}{n\pi} \left( (-1)^{n+1} + \cos\left(\frac{n\pi}{3}\right) \right). \end{aligned}$$

Consequently, the Fourier series of the function at hand has the form  $F(x) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1} + \cos(\frac{n\pi}{3})}{n} \sin(nx)$ .

(3) To confirm the previous expression in Sage it suffices to confirm the expressions of the Fourier coefficients, which can be done via the method presented in 7.A.10, that is,

```
f1(x)=-1; f2(x)=0; f3(x)=1
g=piecewise([[(-pi, -pi/3), f1], [(-pi/3, pi/3), f2], [(pi/3, pi), f3]]); L=pi; n=var("n")
an=(1/L)*integral(f1(x)*cos(n*pi*x/L), x, -pi, -pi/3)+(1/L)*integral(f2(x)*cos(n*pi*x/L), x, -pi/3, pi/3)\
+(1/L)*integral(f3(x)*cos(n*pi*x/L), x, pi/3, pi); show("a_n=", an.expand())
bn=(1/L)*integral(f1(x)*sin(n*pi*x/L), x, -pi, -pi/3)+(1/L)*integral(f2(x)*sin(n*pi*x/L), x, -pi/3, pi/3)\
+(1/L)*integral(f3(x)*sin(n*pi*x/L), x, pi/3, pi); show("b_n=", bn.expand())
```

Executing this block, we obtained the desired result:

$$a_n=0, \quad b_n = -\frac{2 \cos(\pi n)}{\pi n} + \frac{2 \cos\left(\frac{1}{3} \pi n\right)}{\pi n}.$$

(4) An application of the given formula in (2) gives

$$F_5(x) = \frac{3 \sin(x)}{\pi} - \frac{3 \sin(2x)}{2\pi} - \frac{3 \sin(4x)}{4\pi} + \frac{3 \sin(5x)}{5\pi}.$$

To confirm this expression in Sage, simply add the following cell to the block above:

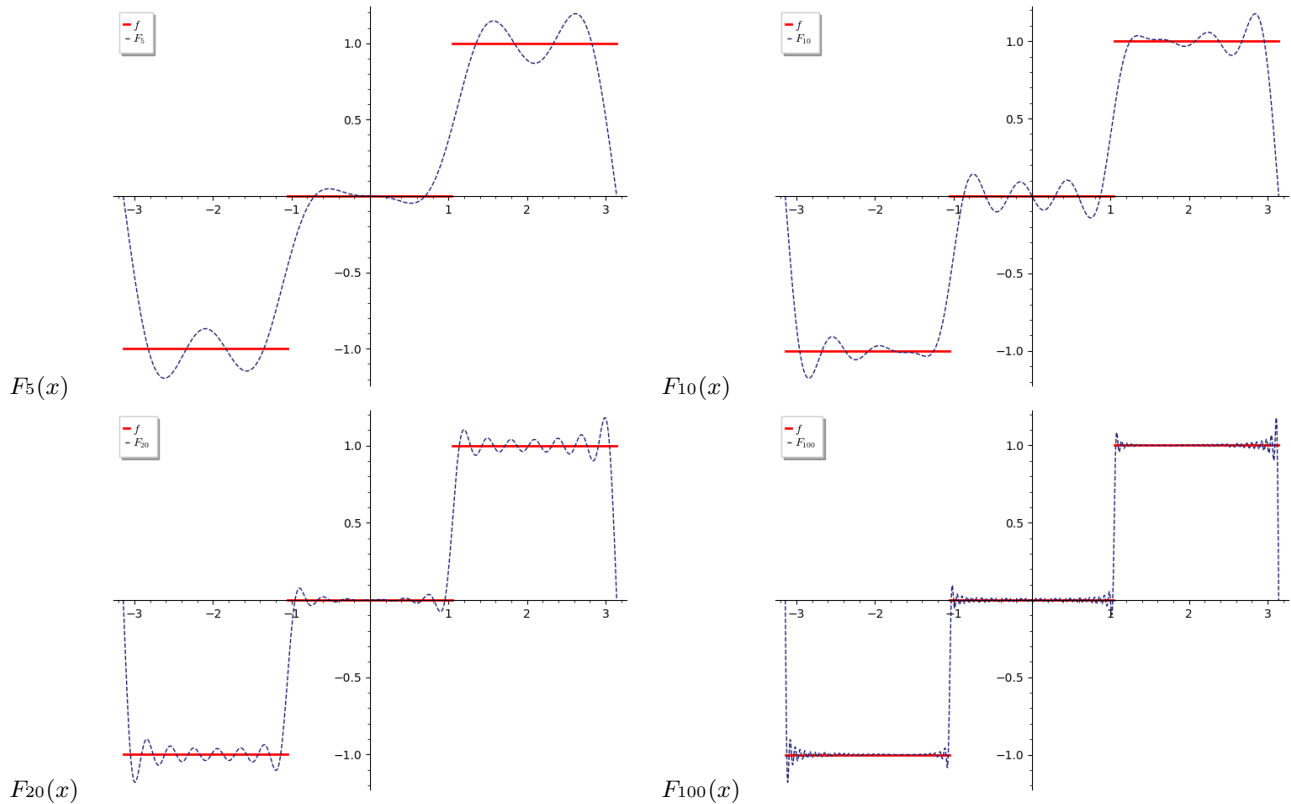
```
pS5 = f.fourier_series_partial_sum(5,pi); show("5th partial sum = ",pS5)
```

To generate the required graphs, include the following syntax in the initial block in (3):

```
pS5 = f.fourier_series_partial_sum(5,pi); pS10 = f.fourier_series_partial_sum(10,pi)
pS20 = f.fourier_series_partial_sum(20,pi); pS100 = f.fourier_series_partial_sum(100,pi)
a = f.plot(x, -pi, pi, thickness=2, color="red", legend_label=r"$f$", exclude=[-pi/3, pi/3])
```

```
s1 = plot(pS5, x, -pi, pi, linestyle="--", color="midnightblue", legend_label=r"$F_5$")
s2 = plot(pS10, x, -pi, pi, linestyle="--", color="midnightblue", legend_label=r"$F_{10}$")
s3 = plot(pS20, x, -pi, pi, linestyle="--", color="midnightblue", legend_label=r"$F_{15}$")
s4 = plot(pS100, x, -pi, pi, linestyle="--", color="midnightblue", legend_label=r"$F_{100}$")
(a+s1).show(); (a+s2).show(); (a+s3).show(); (a+s4).show()
```

Running the code all together, we obtain the following figures:



(5) By the result in (4) we already know the 5th coefficient which is given by  $\frac{3}{5\pi}$ . For a Sage description of the 10th and 100th coefficient of  $F$ , one can proceed as described in 7.A.14. In fact, the Fourier series is a sine series, hence we need the function `f.fourier_series_sine_coefficient(k)`, and the implementation of the method is as follows:

```
f1(x)=-1; f2(x)=0; f3(x)=1
f=piecewise([[(-pi, -pi/3), f1], [[-pi/3, pi/3], f2],[[pi/3, pi), f3]])
Fc5=f.fourier_series_sine_coefficient(5)
show("5th coefficient = ",Fc5)
Fc10=f.fourier_series_sine_coefficient(10)
show("10th coefficient = ",Fc10)
Fc100=f.fourier_series_sine_coefficient(100)
show("100th coefficient = ",Fc100)
```

This confirms the value of the 5th coefficient and additionally provides the values of the other two coefficients in question, namely:

$$10\text{th coefficient} = -\frac{3}{10\pi}, \quad 100\text{th coefficient} = -\frac{3}{100\pi}.$$

Can you guess which coefficient is the 1000th? □

**7.D.9. (Reverse) sawtooth function.** Consider the function defined by

$$f(x) = \begin{cases} \frac{\pi - x}{2}, & \text{if } 0 < x \leq 2\pi, \\ f(x + 2\pi), & \text{otherwise.} \end{cases}$$

- (1) Describe the given function  $f$  for  $x \in (-2\pi, 4\pi]$ , and next use Sage to plot its graph over this interval.
- (2) Find the Fourier series of  $f$  and next use Sage to confirm the result.

(3) Use Sage to plot the partial sums  $F_5$ ,  $F_8$ ,  $F_{15}$  and  $F_{50}$ , in one single figure.

(4) Examine the convergence of the Fourier series and show that

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

**Solution.** (1) The given function  $f$  provides the  $2\pi$ -periodic extension of the function  $(\pi - x)/2$ , the latter defined on  $(0, 2\pi]$ .

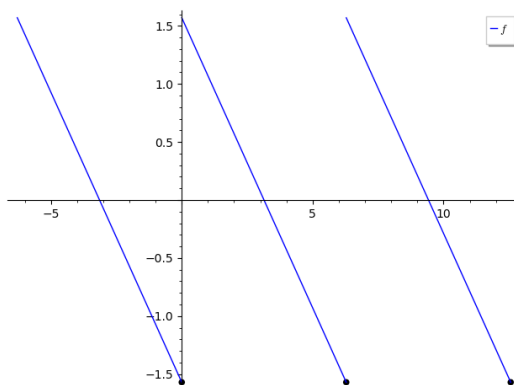
Hence, it should be  $f(x) = f(x + 2\pi) = \frac{\pi - x - 2\pi}{2} = \frac{-\pi - x}{2}$ , for all  $x \in (-2\pi, 0]$  and  $f(x) = f(x - 2\pi) = \frac{\pi - x + 2\pi}{2} = \frac{3\pi - x}{2}$ , for all  $x \in (2\pi, 4\pi]$ , that is,

$$f(x) = \begin{cases} \frac{-\pi - x}{2}, & \text{if } -2\pi < x \leq 0, \\ \frac{\pi - x}{2}, & \text{if } 0 < x \leq 2\pi, \\ \frac{3\pi - x}{2}, & \text{if } 2\pi < x \leq 4\pi. \end{cases}$$

To sketch the graph of this piecewise function use the `piecewise` command, as before. For clarity, you may also mark the endpoints where  $f$  is defined (e.g., with black dots). This can be done using the command `point`, as follows:

```
f1(x)= (pi-x)/2;f2(x)= (-pi-x)/2;f3(x)= (3*pi-x)/2
f = piecewise([[(-2*pi, 0),f2], [(0, 2*pi),f1], [(2*pi, 4*pi), f3]])
p=f.plot(x, -2*pi, 4*pi, exclude=[0, 2*pi], legend_label=r"$f$")
p0=point([0, -pi/2], size=30, color="black")
p1=point([2*pi, -pi/2], size=30, color="black")
p2=point([4*pi, -pi/2], size=30, color="black")
show(p+p0+p1+p2)
```

Below one can see the produced plot:



(2) By assumption the period is  $2\pi$ , so the Fourier series has the form  $F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + b_n \sin(nx))$ , where

$$a_0 = \frac{1}{\pi} \int_0^{2\pi} f(x) dx, \quad a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(nx) dx, \quad b_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(nx) dx.$$

First we will prove that  $a_n = 0$  for all  $n \in \mathbb{N}$ . It is very easy to see that  $a_0 = 0$ . Next, based on integration by parts, we get

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(nx) dx = \frac{1}{2\pi} \int_0^{2\pi} (\pi - x) \cos(nx) dx = \frac{1}{2} \int_0^{2\pi} \cos(nx) dx - \frac{1}{2\pi} \int_0^{2\pi} x \cos(nx) dx \\ &= \frac{1}{2} \left[ \frac{\sin(nx)}{n} \right]_0^{2\pi} - \frac{1}{2\pi} \int_0^{2\pi} x \cdot \left( \frac{\sin(nx)}{n} \right)' dx \\ &= \frac{1}{2n} (\sin(2n\pi) - \sin(0)) - \frac{1}{2\pi} \left( \left[ x \frac{\sin(nx)}{n} \right]_0^{2\pi} - \frac{1}{n} \int_0^{2\pi} \sin(nx) dx \right) \\ &= 0 - \frac{1}{2n\pi} (2\pi \sin(2n\pi) - 0) + \frac{1}{2n\pi} \left[ -\frac{\cos(nx)}{n} \right]_0^{2\pi} = 0 - 0 - \frac{1}{2n^2\pi} (\cos(2n\pi) - \cos(0)) = 0, \end{aligned}$$

for all positive integers  $n$ . In a similar way, for  $b_n$  we compute

$$\begin{aligned}
 b_n &= \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(nx) dx = \frac{1}{2\pi} \int_0^{2\pi} (\pi - x) \sin(nx) dx = \frac{1}{2} \int_0^{2\pi} \sin(nx) dx - \frac{1}{2\pi} \int_0^{2\pi} x \sin(nx) dx \\
 &= -\frac{1}{2n} [\cos(nx)]_0^{2\pi} + \frac{1}{2\pi} \int_0^{2\pi} x \left( \frac{\cos(nx)}{n} \right)' dx \\
 &= -\frac{1}{2n} (\cos(2n\pi) - \cos(0)) + \frac{1}{2\pi} \left( \left[ x \frac{\cos(nx)}{n} \right]_0^{2\pi} - \frac{1}{n} \int_0^{2\pi} \cos(nx) dx \right) \\
 &= -\frac{1}{2n} (1 - 1) + \frac{1}{2n\pi} (2\pi \cos(2n\pi) - 0) - \frac{1}{2n\pi} \left[ \frac{\sin(nx)}{n} \right]_0^{2\pi} = \frac{2\pi}{2n\pi} - \frac{1}{2n^2\pi} (\sin(2n\pi) - \sin(0)) = \frac{1}{n}.
 \end{aligned}$$

Thus the Fourier series associated to  $f$  is given by  $F(x) = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n}$ . To confirm this expression by Sage one can compute the Fourier coefficients, as usual, i.e.,

```

var("n"); f(x)=(pi-x)/2; f = piecewise([(0, 2*pi), f])
an=(1/pi)*integral(((pi-x)/2)*cos(n*x), x, 0, 2*pi)
bn=(1/pi)*integral(((pi-x)/2)*sin(n*x), x, 0, 2*pi)
show("a_n=", an.expand()); show("b_n=", bn.expand())
    
```

(3) Recall that for the partial sums Sage provides the built-in function `fourier_series_partial_sum`. For our case the implementation can be done by continue typing in the previous block the following:

```

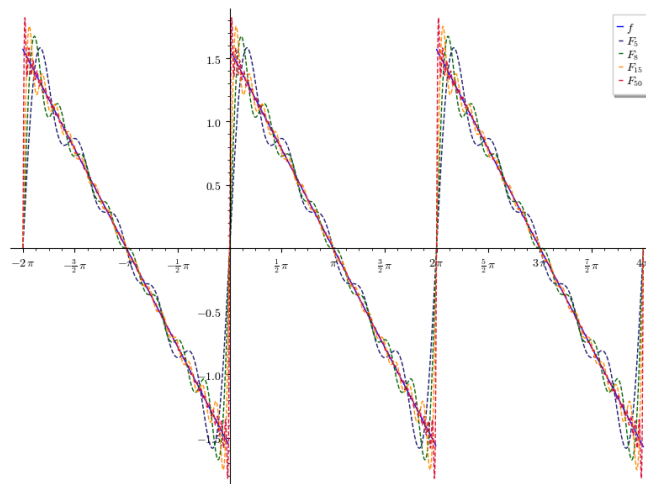
partS5=f.fourier_series_partial_sum(5,pi); show("5th partial sum = ",partS5)
partS8=f.fourier_series_partial_sum(8,pi); show("8th partial sum = ",partS8)
partS15=f.fourier_series_partial_sum(15,pi); show("15th partial sum = ",partS15)
partS50=f.fourier_series_partial_sum(50,pi); show("50th partial sum = ",partS50)
    
```

Run the whole syntax yourselves to obtain the corresponding expressions of the partial sums  $F_5, F_8, F_{15}$  and  $F_{50}$ . To plot them, execute the block given below (notice that this is independent of the previous cells).

```

f1(x)= (pi-x)/2; f2(x)= (-pi-x)/2; f3(x)= (3*pi-x)/2
f = piecewise([(-2*pi, 0),f2], [(0, 2*pi),f1], [(2*pi, 4*pi), f3])
partS5=f.fourier_series_partial_sum(5,pi);partS8=f.fourier_series_partial_sum(8,pi)
partS15=f.fourier_series_partial_sum(15,pi);partS50=f.fourier_series_partial_sum(50,pi)
p=f.plot(x, -2*pi, 4*pi, exclude=[0, 2*pi], legend_label=r"$f$")
s5=plot(partS5, x, -2*pi, 4*pi, linestyle="--", color="midnightblue", legend_label=r"$F_5$")
s8=plot(partS8, x, -2*pi, 4*pi, linestyle="--", color="darkgreen", legend_label=r"$F_8$")
s15=plot(partS15, x, -2*pi, 4*pi, linestyle="--", color="darkorange", legend_label=r"$F_{15}$")
s50=plot(partS50, x, -2*pi, 4*pi, linestyle="--", color="crimson", legend_label=r"$F_{50}$")
(p+s5+s8+s15+s50).show(figsize=8, ticks=pi/2, tick_formatter=pi)
    
```

This generates the following figure:





We should mention that in this block  $f$  was introduced as in the first part (see (1)), with aim to display a larger portion of the graph of  $f$ . To verify the computations in part (2), however, it was adequate to use only the function  $(\pi - x)/2$  on its initial domain.

(4) By the figure above, it is clear that the Fourier series  $F(x)$  of  $f$  converges to  $f(x)$ , at each point  $x$  where  $f$  is continuous, which means that for any  $x \in (0, 2\pi)$  we have

$$\frac{\pi - x}{2} = F(x) = \sum_{n=1}^{\infty} \frac{\sin(nx)}{n} = \sin(x) + \frac{\sin(2x)}{2} + \frac{\sin(3x)}{3} + \frac{\sin(4x)}{4} + \frac{\sin(5x)}{5} + \dots$$

This applies to  $x = \pi/2$  and gives

$$\frac{\pi}{4} = \sin\left(\frac{\pi}{2}\right) + \frac{\sin(\pi)}{2} + \frac{\sin(3\pi/2)}{3} + \frac{\sin(2\pi)}{4} + \frac{\sin(5\pi/2)}{5} + \dots = 1 + 0 - \frac{1}{3} + 0 + \frac{1}{5} - \dots$$

On the other hand, the points of discontinuity of  $f$  appear at  $x = 2k\pi$ , for  $k \in \mathbb{Z}$ . The average value of our function at these points equals to 0, and hence the Fourier expansion of  $f$  converges to 0 at any jump discontinuity of  $f$ .  $\square$

**7.D.10.** Compute the complex version of the Fourier series of the  $2\pi$ -periodic extension of the following function (often referred to as the “pulse wave”)

$$g(x) = \begin{cases} 0, & \text{if } -\pi < x < 0, \\ 1, & \text{if } 0 < x < \pi. \end{cases}$$

$\circ$

**7.D.11.** Consider the exponential mapping  $f(x) = e^x$  with  $x \in [0, 1)$ . Use Sage to implement the following:

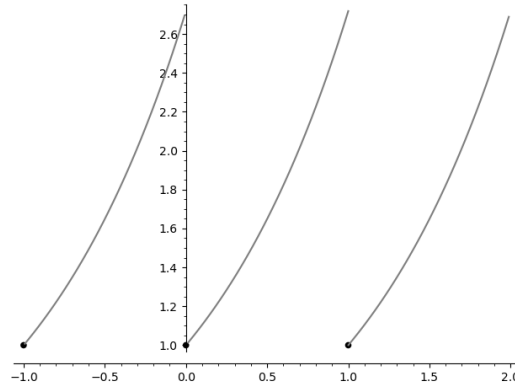
- (1) Determine and illustrate the 1-periodic extension of  $f$  for  $x \in [-1, 2)$  (with at least two different Sage methods);
- (2) Compute the Fourier series of  $f$  on the interval  $[0, 1)$ , and then illustrate its 5th, 15th and 25th partial sum (with at least two different Sage methods);
- (3) Determine the convergence of the Fourier series obtained above.

**Solution.** (1) The 1-periodic extension of  $f$  on the interval  $[-1, 2)$  is given by  $f(x + 1)$  for  $-1 \leq x < 0$ , by  $f(x)$  for  $0 \leq x < 1$  and  $f(x - 1)$  for  $1 \leq x < 2$ . In Sage one method utilizes the `piecewise` function while another one relies on the `def` procedure to define piecewise functions. The implementation goes as follows:

```
f=piecewise([[RealSet.closed_open(-1, 0), e^(x+1)], [(0, 1), e^x], [(1, 2), e^(x-1)]])
a=f.plot(x, -1, 2, thickness=1.5, color="grey", exclude=[0, 1])
p0=point([-1, 1], size=30, color="black")
p1=point([0, 1], size=30, color="black")
p2=point([1, 1], size=30, color="black")
show(a+p0+p1+p2)
```

```
var("t")
def g(t):
    if 0 <= t < 1:
        return e^t
    elif -1 <= t < 0:
        return e^(t+1)
    elif 1 <= t < 2:
        return e^(t-1)
b=plot(g, (-1, 2), exclude=[0, 1], thickness=1.5, color="grey")
p0=point([-1, 1], size=30, color="black")
p1=point([0, 1], size=30, color="black")
p2=point([1, 1], size=30, color="black")
show(b+p0+p1+p2)
```

They both produce the same figure, which is presented here:



Notice that as in 7.D.9, both blocks include code for marking with black dots the endpoints where the periodic extension is defined.

(2) As usual, it is sufficient to compute the Fourier coefficients. However, one should carefully apply the general formulas from 7.1.6 and here is a possible implementation:

```
f=piecewise([[ (0, 1), e^x ]]); T=1; n=var("n"); Om=pi
an=(2/T)*integral((e^x)*cos(2*n*Om*x), x, 0, 1); show("a_n=", an.expand())
bn=(2/T)*integral((e^x)*sin(2*n*Om*x), x, 0, 1); show("b_n=", bn.expand())
```

Sage's answer has the form

$$a_n = \frac{4\pi n e \sin(2\pi n)}{4\pi^2 n^2 + 1} + \frac{2 \cos(2\pi n) e}{4\pi^2 n^2 + 1} - \frac{2}{4\pi^2 n^2 + 1}, \quad b_n = -\frac{4\pi n \cos(2\pi n) e}{4\pi^2 n^2 + 1} + \frac{4\pi n}{4\pi^2 n^2 + 1} + \frac{2 e \sin(2\pi n)}{4\pi^2 n^2 + 1}.$$

Recalling that  $\sin(2n\pi) = 0$  and  $\cos(2n\pi) = 1$  for all  $n$ , this means that

$$a_n = \frac{2(e-1)}{1+4n^2\pi^2}, \quad (n \in \mathbb{N}), \quad \text{and} \quad b_n = \frac{4n\pi(1-e)}{1+4n^2\pi^2}, \quad (n \in \mathbb{Z}_+),$$

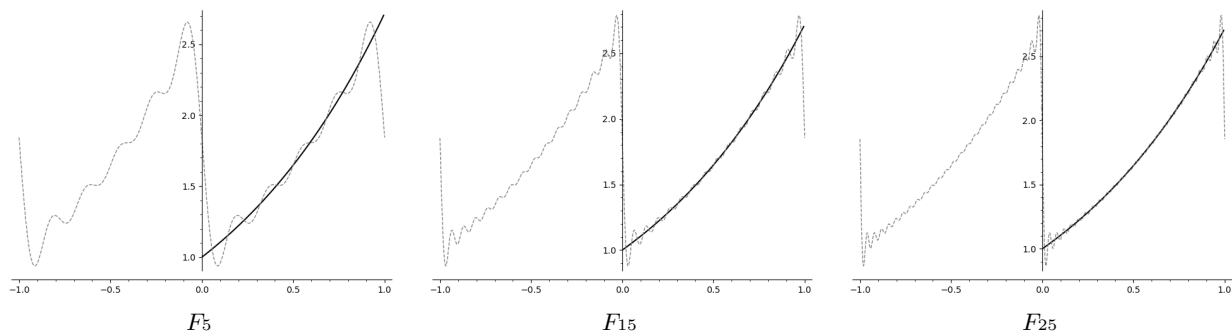
respectively. It follows that the Fourier series of the function at hand has the form

$$F(x) = e - 1 + 2(e-1) \sum_{n=1}^{\infty} \frac{\cos(2n\pi x)}{1+4n^2\pi^2} + 4\pi(1-e) \sum_{n=1}^{\infty} \frac{n \sin(2n\pi x)}{1+4n^2\pi^2}. \quad (b)$$

Now, to plot the required partial sums, we can proceed as in the previous tasks (cf. 7.D.9), i.e.,

```
f=piecewise([[ (0, 1), e^x ]])
partS5 = f.fourier_series_partial_sum(5, 1/2)
partS15 = f.fourier_series_partial_sum(15, 1/2)
partS25 = f.fourier_series_partial_sum(25, 1/2)
a = f.plot(x, 0, 1, thickness=1.5, color="black")
s5=plot(partS5, x, -1, 1, linestyle="--", color="grey")
s15=plot(partS15, x, -1, 1, linestyle="--", color="grey")
s25=plot(partS25, x, -1, 1, linestyle="--", color="grey")
show(a+s5); show(a+s15); show(a+s25)
```

Notice in the first line of this block we could replace the open set  $(0, 1)$  with the correct set that  $f$  is defined, i.e., the set  $[0, 1)$ . This could be done by the command `RealSet.closed_open(0, 1)`, but it is *not* so essential in the sense that it does not effect the Fourier series, as previously mentioned in ???. This block produces the following three figures:



An alternative method to derive the partial sums utilizes a combination of certain built-in functions in Sage, namely `fourier_series_sine_coefficient` and `fourier_series_cosine_coefficient`. Below we present the case for the 5th partial sum and similarly are treated the rest cases.

```
f=piecewise([[0, 1), e**x]])
a5=[f.fourier_series_cosine_coefficient(n, 1/2) for n in range(6)]
b5=[f.fourier_series_sine_coefficient(n, 1/2) for n in range(6)]
fs5=(a5[0]/2)+(sum([a5[k]*cos(2*k*pi*x) for k in range(1, 6)]))
+(sum([b5[k]*sin(2*k*pi*x) for k in range(1, 6)]))
print(bool(fs5.expand()==f.fourier_series_partial_sum(5,1/2).expand()))
a=f.plot(x, 0, 1, thickness=1.5, color="black")
b=fs5.plot(x, 0, 1, linestyle="--")
(a+b).show(figsize=8)
```

In this block we programmed Sage to compare the expression of the 5th partial sum, denoted by `fs5` and defined manually, with the one that one derives via the usual method by the command `fourier_series_partial_sum`. This comparison is performed by the `bool` command, and Sage's answer is `True`. One could also obtain the explicit form of the partial sums in question, which we skipped to save some space.

As a side remark, we finally mention that for a formal verification of the Fourier coefficients  $a_n, b_n$  one can use the following formulas

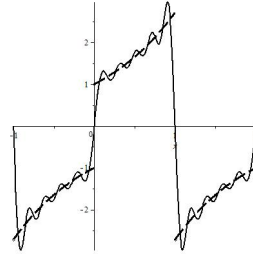
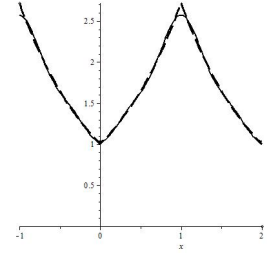
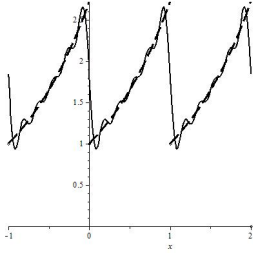
$$\int e^x \cos(\alpha x) dx = \frac{e^x (\alpha \sin(\alpha x) + \cos(\alpha x))}{1 + \alpha^2} + C, \quad \int e^x \sin(\beta x) dx = \frac{e^x (\sin(\beta x) - \beta \cos(\beta x))}{1 + \beta^2} + C,$$

with  $\alpha, \beta \in \mathbb{R}$ . Both of them can be derived by integrations by parts (see 6.B.7 for the second one).

(3) The function  $f(x) = e^x$  is continuous at any  $x \in (0, 1)$ . It is also continuous at  $x = 0$  since  $\lim_{x \rightarrow 0^-} e^x = 1 = \lim_{x \rightarrow 0^+} f(x)$ . It also satisfies according to Dirichlet's theorem (see ??) we should have  $e^x = F(x)$  for all  $x \in [0, 1]$ , where  $F$  is given in (b).  $\square$

- 7.D.12.** (1) Derive a cosine Fourier series of  $f$  on the interval  $[0, 1]$  and illustrate some of its partial sums;  
 (2) Derive a sine Fourier series of  $f$  on the interval  $(0, 1]$  and illustrate some of its partial sums;

Note the differences between the three cases as shown in the diagrams. The approximation differs in relation to the continuity. The first diagram uses  $n = 5$ , the second uses  $n = 3$ , while the last diagram uses  $n = 10$ .



We use the formulae ( $\alpha, \beta \in \mathbb{R}$ )

$$(1) \int e^x \cos(\alpha x) dx = \frac{e^x (\alpha \sin(\alpha x) + \cos(\alpha x))}{1 + \alpha^2} + C,$$

$$(2) \int e^x \sin(\beta x) dx = \frac{e^x (\sin(\beta x) - \beta \cos(\beta x))}{1 + \beta^2} + C,$$

both of which can be obtained by two integrations by parts. Actually, the second one was computed in detail in ??(d). We obtain

$$\begin{aligned} (a) \quad a_n &= 2 \int_0^1 e^x \cos(2n\pi x) dx \\ &= 2 \left[ \frac{e^x (2n\pi \sin(2n\pi x) + \cos(2n\pi x))}{1 + 4n^2 \pi^2} \right]_0^1 \\ &= \frac{2(e-1)}{1 + 4n^2 \pi^2}, \quad n \in \mathbb{N} \cup \{0\}, \end{aligned}$$

$$\begin{aligned} b_n &= 2 \int_0^1 e^x \sin(2n\pi x) dx \\ &= 2 \left[ \frac{e^x (\sin(2n\pi x) - 2n\pi \cos(2n\pi x))}{1 + 4n^2 \pi^2} \right]_0^1 \\ &= \frac{4n\pi(1-e)}{1 + 4n^2 \pi^2}, \quad n \in \mathbb{N}; \end{aligned}$$

$$\begin{aligned} (b) \quad a_n &= 2 \int_0^1 e^x \cos(n\pi x) dx \\ &= 2 \left[ \frac{e^x (n\pi \sin(n\pi x) + \cos(n\pi x))}{1 + n^2 \pi^2} \right]_0^1 \\ &= \frac{2((-1)^n e - 1)}{1 + n^2 \pi^2}, \quad n \in \mathbb{N} \cup \{0\}; \end{aligned}$$

$$\begin{aligned} (c) \quad b_n &= 2 \int_0^1 e^x \sin(n\pi x) dx \\ &= 2 \left[ \frac{e^x (\sin(n\pi x) - n\pi \cos(n\pi x))}{1 + n^2 \pi^2} \right]_0^1 \\ &= \frac{2n\pi(1 + (-1)^{n+1} e)}{1 + n^2 \pi^2}, \quad n \in \mathbb{N}. \end{aligned}$$

Substitution then yields the corresponding Fourier series for  $g(x)$ :

$$\begin{aligned} (a) \quad & e - 1 + 2(e - 1) \sum_{n=1}^{\infty} \frac{\cos(2n\pi x)}{1 + 4n^2 \pi^2} \\ & + 4\pi(1 - e) \sum_{n=1}^{\infty} \frac{n \sin(2n\pi x)}{1 + 4n^2 \pi^2}; \end{aligned}$$

(b)  $e - 1 + 2 \sum_{n=1}^{\infty} \frac{((-1)^n e - 1) \cos(n\pi x)}{1 + n^2 \pi^2}$ ;  
 (c)  $2\pi \sum_{n=1}^{\infty} \frac{n(1 + (-1)^{n+1} e) \sin(n\pi x)}{1 + n^2 \pi^2}$ .

**7.D.13.** Determine the cosine Fourier series for a periodic extension of the function

$$g(x) = 1, \quad x \in [0, 1), \quad g(x) = 0, \quad x \in [1, 4).$$

Determine also the sine Fourier series for

$$f(x) = x - 1, \quad x \in (0, 2), \quad f(x) = 3 - x, \quad x \in [2, 4).$$

**Solution.** We have already encountered the construction of a cosine Fourier series in ??(b) and also 7.A.15(b). It is the case of the Fourier series of an even function. Therefore, the first thing we do is extend the definition of the function  $g$  to the interval  $(-4, 0)$  so that it is even. This means

$$g(x) := 1 \text{ for } x \in (-1, 0), \quad g(x) := 0 \text{ for } x \in (-4, -1].$$

Now we can consider its periodic extension onto the whole  $\mathbb{R}$  with period  $x_0 = -4, T = 8$  and  $\omega = \pi/4$ .

Necessarily  $b_n = 0$  for all  $n \in \mathbb{N}$  in a cosine Fourier series. We determine the Fourier coefficients  $a_n, n \in \mathbb{N}$

$$\begin{aligned} a_0 &= \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \, dx = \frac{2}{8} \int_{-1}^1 1 \, dx = \frac{1}{2} \int_0^1 1 \, dx = \frac{1}{2}, \\ a_n &= \frac{2}{T} \int_{x_0}^{x_0+T} g(x) \cos(n\omega x) \, dx = \frac{1}{2} \int_0^1 \cos \frac{n\pi x}{4} \, dx \\ &= \frac{2}{n\pi} \sin \frac{n\pi}{4}. \end{aligned}$$

We use

$$(1) \quad \int_{-a}^a f(x) \, dx = 2 \int_0^a f(x) \, dx,$$

which is valid for every even function  $f$  integrable on the interval  $[0, a]$ .

The expression  $\sin(n\pi/4)$  is conveniently left as is, rather than the alternative of sorting out when it attains which of its five different values. Thus we write the cosine Fourier series in the form:

$$\frac{1}{4} + \sum_{n=1}^{\infty} \left( \frac{2}{n\pi} \sin \frac{n\pi}{4} \cos \frac{n\pi x}{4} \right).$$

The sine Fourier transform of the function can be determined analogously from the odd extension of the given segment. Again,  $T = 8$  and  $\omega = \pi/4$  for the function  $f$ . This time it is the coefficients  $a_n, n \in \mathbb{N} \cup \{0\}$ , which are zero. To find the remaining coefficients, use integration by parts and (1) (the product of two odd functions is an even function). For all  $n \in \mathbb{N}$

$$\begin{aligned} b_n &= \frac{2}{T} \int_{x_0}^{x_0+T} f(x) \sin(n\omega x) \, dx \\ &= \frac{1}{2} \left( \int_0^2 (x-1) \sin \frac{n\pi x}{4} \, dx - \int_2^4 (x-3) \sin \frac{n\pi x}{4} \, dx \right) \\ &= \left[ (1-x) \frac{2}{n\pi} \cos \frac{n\pi x}{4} \right]_0^2 + \left[ \frac{8}{n^2 \pi^2} \sin \frac{n\pi x}{4} \right]_0^2 \\ &\quad - \left[ (3-x) \frac{2}{n\pi} \cos \frac{n\pi x}{4} \right]_2^4 - \left[ \frac{8}{n^2 \pi^2} \sin \frac{n\pi x}{4} \right]_2^4 \\ &= \frac{2}{n\pi} ((-1)^n - 1) + \frac{16}{n^2 \pi^2} \sin \frac{n\pi}{2}. \end{aligned}$$

Immediately,  $b_n = 0$  when  $n$  is even. So the sine Fourier series can be written

$$\begin{aligned} &\sum_{n=1}^{\infty} \left( \left( \frac{2}{n\pi} ((-1)^n - 1) + \frac{16}{n^2 \pi^2} \sin \frac{n\pi}{2} \right) \sin \frac{n\pi x}{4} \right) \\ &= \sum_{n=1}^{\infty} \left( \left( \frac{-4}{(2n-1)\pi} + \frac{(-1)^{n-1} 16}{(2n-1)^2 \pi^2} \right) \sin \frac{(2n-1)\pi x}{4} \right). \end{aligned}$$

□

**7.D.14.** Write the Fourier series of the  $\pi$ -periodic function which equals cosine on the interval  $(-\pi/2, \pi/2)$ . Then write the cosine Fourier series of the  $2\pi$ -periodic function  $y = |\cos x|$ .

**Solution.** We are looking for one Fourier series only, since the second part of the problem is just a reformulation of the first part. Therefore, we construct the Fourier series for the function  $g(x) = \cos x, x \in [-\pi/2, \pi/2]$ . Since  $g$  is even,  $b_n = 0, n \in \mathbb{N}$ . We compute

$$\begin{aligned}
 a_n &= \frac{2}{\pi} \int_{-\pi/2}^{\pi/2} \cos x \cos(2nx) \, dx \\
 &= \frac{2}{\pi} \int_{-\pi/2}^{\pi/2} \frac{1}{2} (\cos((2n+1)x) + \cos((2n-1)x)) \, dx \\
 &= \frac{1}{\pi} \left[ \frac{\sin((2n+1)x)}{2n+1} + \frac{\sin((2n-1)x)}{2n-1} \right]_{-\pi/2}^{\pi/2} \\
 &= \frac{2}{\pi} \left( \frac{(-1)^n}{2n+1} + \frac{(-1)^{n+1}}{2n-1} \right) \\
 &= \frac{4}{\pi} \frac{(-1)^{n+1}}{4n^2-1}
 \end{aligned}$$

for every  $n \in \mathbb{N}$ . The calculation is also valid for  $n = 0$ , thus  $a_0 = 4/\pi$ . The desired Fourier series is

$$\frac{2}{\pi} + \frac{4}{\pi} \sum_{n=1}^{\infty} \left( \frac{(-1)^{n+1}}{4n^2-1} \cos(2nx) \right).$$

□

**7.D.15.** Sum the two series

$$\sum_{n=1}^{\infty} \frac{1}{n^4}, \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^4}.$$

**Solution.** We hint at the procedure by which the series

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}}, \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^{2k}}$$

for general  $k \in \mathbb{N}$  can be calculated. Use the identities

$$(1) \quad x = \pi - 2 \sum_{n=1}^{\infty} \frac{\sin(nx)}{n}, \quad x \in (0, 2\pi),$$

$$(2) \quad x^2 = \frac{4\pi^2}{3} + 4 \sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2} - 4\pi \sum_{n=1}^{\infty} \frac{\sin(nx)}{n}, \quad x \in (0, 2\pi),$$

which follow from the constructions of the Fourier series for the functions  $g(x) = x$  and  $g(x) = x^2$ , respectively, on the interval  $[0, 2\pi)$ .

By (1),

$$\sum_{n=1}^{\infty} \frac{\sin(nx)}{n} = \frac{\pi-x}{2}, \quad x \in (0, 2\pi).$$

Substituting into (2) gives

$$\sum_{n=1}^{\infty} \frac{\cos(nx)}{n^2} = \frac{3x^2 - 6\pi x + 2\pi^2}{12}, \quad x \in (0, 2\pi).$$

Since the values of the series

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12}$$

have been already determined, substitution then proves the validity of this last equation at the marginal points  $x = 0$ ,  $x = 2\pi$ . The left-hand series is evidently bounded from above by  $\sum_{n=1}^{\infty} \frac{1}{n^2}$ , thus it converges absolutely and uniformly on  $[0, 2\pi]$ . Therefore, it can be integrated term by term:

$$\begin{aligned}
 \sum_{n=1}^{\infty} \frac{\sin(nx)}{n^3} &= \sum_{n=1}^{\infty} \left[ \frac{\sin(ny)}{n^3} \right]_0^x = \int_0^x \sum_{n=1}^{\infty} \frac{\cos(ny)}{n^2} \, dy \\
 &= \int_0^x \frac{3y^2 - 6\pi y + 2\pi^2}{12} \, dy = \frac{x^3 - 3\pi x^2 + 2\pi^2 x}{12}, \quad x \in [0, 2\pi].
 \end{aligned}$$

In fact, every Fourier series may be integrated term by term. Further integration gives

$$\begin{aligned}
 \sum_{n=1}^{\infty} \frac{1 - \cos(nx)}{n^4} &= \sum_{n=1}^{\infty} \left[ -\frac{\cos(ny)}{n^4} \right]_0^x = \int_0^x \sum_{n=1}^{\infty} \frac{\sin(ny)}{n^3} \, dy \\
 &= \int_0^x \frac{y^3 - 3\pi y^2 + 2\pi^2 y}{12} \, dy = \frac{x^4 - 4\pi x^3 + 4\pi^2 x^2}{48}, \quad x \in [0, 2\pi].
 \end{aligned}$$

Substituting  $x = \pi$  leads to

$$\sum_{n=1}^{\infty} \frac{1+(-1)^{n+1}}{n^4} = \sum_{n=1}^{\infty} \frac{1-\cos(n\pi)}{n^4} = \frac{\pi^4}{48}.$$

Since the numerator on the left-hand side is zero for even numbers  $n$  and is 2 for odd numbers  $n$ , the series can be written

$$(3) \quad \sum_{n=1}^{\infty} \frac{2}{(2n-1)^4} = \frac{\pi^4}{48}.$$

From the expression

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \sum_{n=1}^{\infty} \frac{1}{(2n)^4} + \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \frac{1}{16} \sum_{n=1}^{\infty} \frac{1}{n^4} + \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4},$$

it follows that

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{16}{15} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} = \frac{16}{15} \cdot \frac{1}{2} \cdot \frac{\pi^4}{48} = \frac{\pi^4}{90},$$

thereby having summed up the first series. As for the second one,

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^4} &= \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} - \sum_{n=1}^{\infty} \frac{1}{(2n)^4} = \sum_{n=1}^{\infty} \frac{1}{(2n-1)^4} - \frac{1}{16} \sum_{n=1}^{\infty} \frac{1}{n^4} \\ &= \frac{1}{2} \cdot \frac{\pi^4}{48} - \frac{1}{16} \cdot \frac{\pi^4}{90} = \frac{7\pi^4}{720}. \end{aligned}$$

One can proceed similarly to sum the series

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}}, \quad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^{2k}}$$

for other  $k \in \mathbb{N}$ .

It is natural to ask for the value of the series  $\sum_{n=1}^{\infty} \frac{1}{n^3}$ . This problem has been tackled by mathematicians for centuries without success. The reader may justifiably be surprised by this since the procedure above is applicable to all the odd powers as well.

For instance, one can start with the identity

$$\sum_{n=1}^{\infty} \frac{\cos(nx)}{n} = -\ln \left( 2 \sin \frac{x}{2} \right), \quad x \in (0, 2\pi),$$

which, by the way, can be proved by expanding the right-hand function into a Fourier series. If, similarly to the above, integrate the left-hand series term by term twice and substituted  $x \rightarrow 0+$  in the limit, we get the series  $\sum_{n=1}^{\infty} \frac{1}{n^3}$ . Thus, it should suffice to integrate the right-hand function twice and calculate one limit. However, the integration of the right-hand side leads to a non-elementary integral. That is, the antiderivative cannot be expressed in terms of the elementary functions.<sup>6</sup>

□

**7.D.16.** Determine the function  $f$  whose Fourier transform is the function

$$\tilde{f}(\omega) = \frac{1}{\sqrt{2\pi}} \frac{\sin \omega}{\omega}, \quad \omega \neq 0.$$

**Solution.** We might have noticed that the sinc function appeared as the image of the characteristic function  $h_{\Omega}$  of the interval  $(-\Omega, \Omega)$  in one of the previous problems:

$$\tilde{h}_{\Omega}(\omega) = \frac{2\Omega}{\sqrt{2\pi}} \operatorname{sinc}(\omega\Omega).$$

In this case  $\Omega = 1$  and the function  $f$  is the half of  $h_1$ .

The result can be computed directly. The inverse Fourier transform is

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin \omega}{\omega} e^{i\omega t} d\omega \\ &= \frac{1}{2\pi} \left( \int_{-\infty}^0 \frac{\sin \omega}{\omega} e^{i\omega t} d\omega + \int_0^{\infty} \frac{\sin \omega}{\omega} e^{i\omega t} d\omega \right). \end{aligned}$$

Substitute  $-\omega$  for  $\omega$  in the first integral, to obtain

<sup>6</sup>The function  $\zeta(p) = \sum_{n=1}^{\infty} \frac{1}{n^p}$  is called the Riemann zeta function.  
EXPAND THE FOOTNOTE!

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \left( \int_0^\infty \frac{\sin \omega}{\omega} e^{-i\omega t} d\omega + \int_0^\infty \frac{\sin \omega}{\omega} e^{i\omega t} d\omega \right) \\ &= \frac{1}{2\pi} \int_0^\infty \frac{\sin \omega}{\omega} 2(e^{-i\omega t} + e^{i\omega t}) d\omega = \frac{1}{\pi} \int_0^\infty \frac{\sin \omega}{\omega} \cos(\omega t) d\omega. \end{aligned}$$

Continue via trigonometric identities to obtain

$$f(t) = \frac{1}{2\pi} \left( \int_0^\infty \frac{\sin(\omega(1+t))}{\omega} d\omega + \int_0^\infty \frac{\sin(\omega(1-t))}{\omega} d\omega \right).$$

The substitutions  $u = \omega(1+t)$ ,  $v = \omega(1-t)$  then give

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \left( \int_0^\infty \frac{\sin u}{u} du - \int_0^\infty \frac{\sin v}{v} dv \right) = 0, \quad t > 1; \\ f(t) &= \frac{1}{2\pi} \left( \int_0^\infty \frac{\sin u}{u} du + \int_0^\infty \frac{\sin v}{v} dv \right) = \frac{1}{\pi} \int_0^\infty \frac{\sin u}{u} du, \quad t \in (-1, 1); \\ f(t) &= \frac{1}{2\pi} \left( -\int_0^\infty \frac{\sin u}{u} du + \int_0^\infty \frac{\sin v}{v} dv \right) = 0, \quad t < -1. \end{aligned}$$

Thus the function  $f$  is zero for  $|t| > 1$  and constant (necessarily non-zero) for  $|t| < 1$ . (Throughout, we assume that the inverse Fourier transform exists).

The constant is  $f(t) = 1/2$  for  $|t| < 1$ , from the standard result

$$\int_0^\infty \frac{\sin u}{u} du = \frac{\pi}{2}.$$

Alternatively, we can 'guess' that the constant is one, i.e.

$$g(t) = 1, \quad |t| < 1; \quad g(t) = 0, \quad |t| > 1$$

and compute

$$\mathcal{F}(g)(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-i\omega t} dt = \frac{2}{\sqrt{2\pi}} \int_0^1 \cos(\omega t) dt = \frac{2}{\sqrt{2\pi}} \frac{\sin \omega}{\omega}.$$

So  $f(0) = g(0)/2 = 1/2$ , which also establishes

$$\int_0^\infty \frac{\sin u}{u} du = \frac{\pi}{2}.$$

□

#### 7.D.17. Using the relation

$$(1) \quad \mathcal{L}(f')(s) = s\mathcal{L}(f)(s) - \lim_{t \rightarrow 0^+} f(t),$$

derive the Laplace transforms of both the functions  $y = \cos t$  and  $y = \sin t$ .

**Solution.** Notice first that from (1), it follows that

$$\begin{aligned} \mathcal{L}(f'')(s) &= s\mathcal{L}(f')(s) - \lim_{t \rightarrow 0^+} f'(t) = s \left( s\mathcal{L}(f)(s) - \lim_{t \rightarrow 0^+} f(t) \right) - \lim_{t \rightarrow 0^+} f'(t) = \\ &= s^2\mathcal{L}(f)(s) - s \lim_{t \rightarrow 0^+} f(t) - \lim_{t \rightarrow 0^+} f'(t). \end{aligned}$$

Therefore,

$$-\mathcal{L}(\sin t)(s) = \mathcal{L}(-\sin t)(s) = \mathcal{L}((\sin t)'')(s) = s^2\mathcal{L}(\sin t)(s) - s \lim_{t \rightarrow 0^+} \sin t - \lim_{t \rightarrow 0^+} \cos t = s^2\mathcal{L}(\sin t)(s) - 1,$$

whence we get

$$-\mathcal{L}(\sin t)(s) = s^2\mathcal{L}(\sin t)(s) - 1, \quad \text{i. e.} \quad \mathcal{L}(\sin t)(s) = \frac{1}{s^2+1}.$$

Now, invoking (1), we determine

$$\mathcal{L}(\cos t)(s) = \mathcal{L}((\sin t)')(s) = s \frac{1}{s^2+1} - \lim_{t \rightarrow 0^+} \sin t = \frac{s}{s^2+1}.$$

□



**7.D.18.** Using the discussion from the previous problem, prove that for a continuous function  $y$  with enough sufficiently higher order derivatives

$$\mathcal{L}(y^{(n)})(s) = s^n \mathcal{L}(y)(s) - \sum_{i=1}^n s^{n-i} y^{(i-1)}(0).$$

**Solution.** Clearly

$$\begin{aligned}\mathcal{L}(y')(s) &= s\mathcal{L}(y)(s) - y(0) \\ \mathcal{L}(y'')(s) &= s^2\mathcal{L}(y)(s) - sy(0) - y'(0)\end{aligned}$$

and the claim is verified by induction. □

**7.D.19.** Find the Laplace transform of Heaviside's function  $H(t)$  and, for real  $a$ , the shifted Heaviside's function  $H_a(t) = H(t - a)$ :

$$H(t) = \begin{cases} 0 & \text{for } t < 0, \\ \frac{1}{2} & \text{for } t = 0, \\ 1 & \text{for } t > 0. \end{cases}$$

**Solution.**

$$\begin{aligned}\mathcal{L}(H(t))(s) &= \int_0^\infty H(t) e^{-st} dt = \int_0^\infty e^{-st} dt \\ &= \left[ -\frac{e^{-st}}{s} \right]_0^\infty = -\frac{1}{s}(0 - 1) = \frac{1}{s}, \\ \mathcal{L}(H_a(t))(s) &= \mathcal{L}(H(t - a))(s) \\ &= \int_0^\infty H(t - a) e^{-st} dt = \int_a^\infty e^{-st} dt \\ &= \int_0^\infty e^{-s(t+a)} dt = e^{-as} \mathcal{L}(H(t))(s) = \frac{e^{-as}}{s}.\end{aligned}$$

□

**7.D.20.** Show that for real  $a$ ,

$$(1) \quad \mathcal{L}(f(t) \cdot H_a(t))(s) = e^{-as} \mathcal{L}(f(t + a))(s)$$

**Solution.**

$$\begin{aligned}\mathcal{L}(f(t)H_a(t))(s) &= \int_0^\infty f(t)H(t - a)e^{-st} dt = \int_a^\infty f(t)e^{-st} dt \\ &= \int_0^\infty f(t + a)e^{-s(t+a)} dt = e^{-as} \int_0^\infty f(t + a)e^{-st} dt \\ &= e^{-as} \mathcal{L}(f(t + a))(s).\end{aligned}$$

□

**7.D.21.** Find a function  $y(t)$  satisfying the differential equation

$$y''(t) + 4y(t) = \sin 2t$$

and the initial conditions  $y(0) = 0$  and  $y'(0) = 0$ .

**Solution.** From the example 7.D.18:

$$s^2\mathcal{L}(y)(s) + 4\mathcal{L}(y)(s) = \mathcal{L}(\sin 2t)(s)$$

Now, by 7.B.17(d)

$$\mathcal{L}(\sin 2t)(s) = \frac{2}{s^2 + 4}.$$

It follows that

$$\mathcal{L}(y)(s) = \frac{2}{(s^2 + 4)^2}.$$

The inverse transform then gives

$$y(t) = \frac{1}{8} \sin 2t - \frac{1}{4}t \cos 2t.$$

□

**7.D.22.** Find a function  $y(t)$  satisfying the differential equation and the initial conditions:

$$y''(t) + 4y(t) = f(t), \quad y(0) = 0, \quad y'(0) = -1,$$

where  $f(t)$  is the piecewise continuous function

$$f(t) = \begin{cases} \cos(2t) & \text{for } 0 \leq t < \pi, \\ 0 & \text{for } t \geq \pi. \end{cases}$$

**Solution.** This problem is a model for the undamped oscillations of a spring (excluding friction and other phenomena like non-linearities in the toughness of the spring and so on). It is initiated by an exterior force during the initial period only and then ceases.

The function  $f(t)$  can be written as a linear combination of Heaviside's function  $H(t)$  and its shift. That is,

$$f(t) = \cos(2t)(H(t) - H_\pi(t))$$

up to the values  $t = 0, t = \pi, \dots$  Since

$$\mathcal{L}(y'')(s) = s^2 \mathcal{L}(y) - sy(0) - y'(0) = s^2 \mathcal{L}(y) + 1,$$

we get, making use of the previous example 7.D.20, the right-hand sides to the calculation of the Laplace transform

$$\begin{aligned} s^2 \mathcal{L}(y) + 1 + 4\mathcal{L}(y) &= \mathcal{L}(\cos(2t)(H(t) - H_\pi(t))) \\ &= \mathcal{L}(\cos(2t) \cdot H(t)) - \mathcal{L}(\cos(2t) \cdot H_\pi(t)) \\ &= \mathcal{L}(\cos(2t)) - e^{-\pi s} \mathcal{L}(\cos(2(t + \pi))) \\ &= (1 - e^{-\pi s}) \frac{s}{s^2 + 4}. \end{aligned}$$

Hence,

$$\mathcal{L}(y) = -\frac{1}{s^2 + 4} + (1 - e^{-\pi s}) \frac{s}{(s^2 + 4)^2}.$$

The inverse transform then yields the solution in the form

$$y(t) = -\frac{1}{2} \sin(2t) + \frac{1}{4}t \sin(2t) + \mathcal{L}^{-1} \left( e^{-\pi s} \frac{s}{(s^2 + 4)^2} \right).$$

According to (1),

$$\begin{aligned} \mathcal{L}^{-1} \left( e^{-\pi s} \frac{s}{(s^2 + 4)^2} \right) &= \frac{1}{4} \mathcal{L}^{-1} (e^{-\pi s} \mathcal{L}(t \sin(2t))) \\ &= (t - \pi) \sin(2(t - \pi)) \cdot H_\pi(t). \end{aligned}$$

Since the Heaviside function  $H_\pi(t)$  is zero for  $t < \pi$  and equals 1 for  $t > \pi$ , we get the solution in the form

$$y(t) = \begin{cases} \frac{t-2}{4} \sin(2t) & \text{for } 0 \leq t < \pi \\ \left( \frac{5t-2}{4} - \pi \right) \sin(2t) & \text{for } t \geq \pi. \end{cases}$$

□

More details ?

check the reference

More details!

Key to the exercises

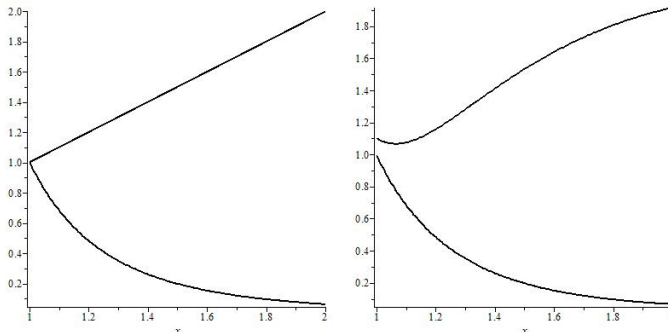
**7.A.2.** Similarly to 7.A.1, one can use the Gram–Schmidt process with respect to the mentioned scalar product. This gives the functions

$$f_1(x) = \frac{1}{x}, \quad f_2(x) = \frac{1}{x^2} - \frac{3}{4x}, \quad f_3(x) = \frac{1}{x^3} - \frac{3}{2x^2} + \frac{13}{24x}.$$

and it is easy to check that the set  $\{f_1, f_2, f_3\}$  forms an orthonormal basis. We also see that:

- The projection of the function  $\frac{1}{x^4}$  has the form  $\frac{15}{32}f_1 + \frac{69}{40}f_2 + \frac{9}{4}f_3$ , while the distance is  $\frac{\sqrt{14}}{2240} = 0.00167$ .
- The projection of the function  $x$  has the form  $2f_1 + 96(-\frac{3}{4} + \ln(2))f_2 + 5760(-\frac{3}{2} \ln(2) + \frac{25}{24})f_3$ , while the distance is approximately 0.035.

Let us now illustrate the functions  $x$  and  $1/x^4$  and their approximations.



In this figure we see that the function  $\frac{1}{x^4}$ , which is the one whose shape is similar to that of one or more generators, is approximated much better by the projection.

**7.A.7.** We have just to check the orthogonality of the couples  $L_m(x), L_n(x)$  with respect to the inner product  $\langle f, g \rangle_\omega = \int_0^\infty f(t)g(t) e^{-t} dt$ . This can be done by integration by parts.

**7.A.8.** The claim follows from the fact that the powers  $x^k$  appear in the polynomials  $T_k$  or  $L_k$  the first time. Thus the linear hulls of the first  $k$  functions always coincide.

**7.A.9.** We mention that since  $f, g$  are both piecewise continuous functions defined on  $[-\pi, \pi]$ , the product  $f\bar{g}$  also belongs to  $\mathcal{S}^0([-\pi, \pi])$ . Thus, the function  $f\bar{g}$  is Riemann integrable and  $(f, g) := \int_{-\pi}^\pi f(x)\bar{g}(x)dx$  is well-defined. The rest details proving that  $(\cdot, \cdot)$  is an inner product on  $\mathcal{S}^0([-\pi, \pi])$  rely on elementary properties of integrals. Next, we see that

$$\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = \frac{1}{\pi} \int_{-\pi}^\pi \frac{1}{2} dx = \frac{1}{\pi} \left[\frac{x}{2}\right]_{-\pi}^\pi = \frac{\pi}{\pi} = 1.$$

Moreover, for any positive integer  $n$  we have

$$\begin{aligned} \left(\frac{\sqrt{2}}{2}, \sin(nx)\right) &= \frac{\sqrt{2}}{2\pi} \int_{-\pi}^\pi \sin(nx) dx = -\frac{\sqrt{2}}{2\pi} \left[\frac{1}{n} \cos(nx)\right]_{-\pi}^\pi = -\frac{\sqrt{2}}{2n\pi} [\cos(n\pi) - \cos(-n\pi)] = 0, \\ \left(\frac{\sqrt{2}}{2}, \cos(nx)\right) &= \frac{\sqrt{2}}{2\pi} \int_{-\pi}^\pi \cos(nx) dx = 0, \\ (\sin(nx), \sin(nx)) &= \frac{1}{\pi} \int_{-\pi}^\pi \sin^2(nx) dx = \frac{1}{2\pi} \int_{-\pi}^\pi (1 - \cos(2nx)) dx = \frac{1}{2\pi}(\pi + \pi) - \frac{1}{2\pi} \left[\frac{1}{2n} \sin(2n)\right]_{-\pi}^\pi = 1, \\ (\cos(nx), \cos(nx)) &= \frac{1}{\pi} \int_{-\pi}^\pi \cos^2(nx) dx = \frac{1}{2\pi} \int_{-\pi}^\pi (1 + \cos(2nx)) dx = 1. \end{aligned}$$

Finally, using appropriate trigonometric identities it is not hard to prove that

$$(\cos(mx), \cos(nx)) = (\cos(mx), \sin(nx)) = (\sin(mx), \sin(nx)) = 0$$

for all positive integers  $m, n$  with  $m \neq n$ . Moreover, we have  $(\cos(mx), \sin(nx)) = 0$  also for  $m = n$ , and the claim follows.

**7.A.12.** A confirmation of the expression presented in 7.A.11 goes as follows:

```
f = piecewise([[(-pi, pi), abs(x)]]); L=pi; n=var("n")
an=(1/L)*integral(abs(x)*cos(n*pi*x/L), x, -pi, pi); show("a_n=", an.expand())
bn=(1/L)*integral(abs(x)*sin(n*pi*x/L), x, -pi, pi); show("b_n=", bn.expand())
```

For the approximation with  $k = 5$ , and its graph, add in the previous block the cell

```
Fs5 = f.fourier_series_partial_sum(5,pi); show("5th partial sum = ",Fs5)
s5 = plot(Fs5, x, -pi-0.5, pi+0.5, linestyle="--", tick_formatter=[2*pi, None])
a=plot(f, x, -pi, pi, color="black", thickness=1.5)
show(a+s5)
```

Execute this code to obtain the explicit form of  $F_5(x)$  and confirm graphically that it provides a pretty good approximation of the function at hand. The confirmation of the identity  $F_5(x) = F_6(x)$  can be easily done and we leave this for practice.

**7.A.13.** The sign function has discontinuities at  $x = 0$  and  $x = \pm\pi$ . Since  $\operatorname{sgn}(-x) = -\operatorname{sgn}(x)$  for all  $x$  we get  $a_n = 0$  for all  $n \in \mathbb{N}$ , and the Fourier series of  $\operatorname{sgn}(x)$  is a sine series,  $F(x) = \sum_{n=1}^{\infty} b_n \sin(nx)$ . The function  $g(x) = \operatorname{sgn}(x) \sin(nx)$  is even, as a product of two odd functions, thus

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} \operatorname{sgn}(x) \sin(nx) dx = \frac{2}{\pi} \int_0^{\pi} \sin(nx) dx = \frac{2}{\pi} \left[ -\frac{\cos(nx)}{n} \right]_0^{\pi} = \frac{2}{\pi} \left[ -\frac{\cos(n\pi)}{n} + \frac{1}{n} \right] = \frac{2(1 - (-1)^n)}{n\pi}.$$

We can simplify this by observing that

$$b_n = \begin{cases} 0, & \text{if } n = 2m, \text{ where } m \in \mathbb{Z}_+, \\ \frac{4}{n\pi}, & \text{if } n = 2m - 1, \text{ where } m \in \mathbb{Z}_+. \end{cases}$$

Therefore, the corresponding Fourier series has the form

$$F(x) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1 - (-1)^n}{n} \sin(nx) = \frac{4}{\pi} \sum_{m=1}^{\infty} \frac{1}{(2m-1)} \sin((2m-1)x).$$

**7.A.17.** Since  $f$  is neither even, nor odd, the complex Fourier coefficients will have both real and imaginary part. In particular, using the properties of the exponential map, we compute

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^x e^{-inx} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{(1-in)x} dx = \frac{1}{2\pi} \left[ \frac{e^{(1-in)x}}{(1-in)} \right]_{-\pi}^{\pi} = \frac{1}{2\pi(1-in)} \left( e^{(1-in)\pi} - e^{-(1-in)\pi} \right).$$

Now it is easy to see that  $e^{(1-in)\pi} = e^{\pi} \cos(n\pi)$  and  $e^{-(1-in)\pi} = e^{-\pi} \cos(n\pi)$ . Thus, recalling the identity  $2 \sinh(x) = (e^x - e^{-x})$ , we have arrived at the following expression:

$$c_n = \frac{\cos(n\pi)}{2\pi(1-in)} (e^{\pi} - e^{-\pi}) = \frac{\sinh(\pi)}{\pi} \frac{(1+in)}{(1+n^2)} \cos(n\pi) = \frac{\sinh(\pi)}{\pi} \frac{(1+in)}{(1+n^2)} (-1)^n.$$

This also gives  $c_0 = \frac{\sinh(\pi)}{\pi} = \frac{e^{\pi} - e^{-\pi}}{2\pi}$ . As for  $c_{-n}$  we see that

$$c_{-n} = \frac{\sinh(\pi)}{\pi} \frac{(1-in)}{(1+(-n)^2)} \cos(-nx) = \frac{\sinh(\pi)}{\pi} \frac{(1-in)}{(1+n^2)} (-1)^n = \overline{c_n},$$

as it is required. Recalling now that  $e^{inx} + e^{-inx} = 2 \cos(nx)$  and  $e^{inx} - e^{-inx} = 2i \sin(nx)$ , we finally get

$$\begin{aligned} F(x) &= c_0 + \sum_{n=1}^{\infty} c_n e^{inx} + \sum_{n=1}^{\infty} c_{-n} e^{-inx} \\ &= \frac{\sinh(\pi)}{\pi} \left( 1 + \sum_{n=1}^{\infty} \frac{(1+in)}{(1+n^2)} (-1)^n e^{inx} + \sum_{n=1}^{\infty} \frac{(1-in)}{(1+n^2)} (-1)^n e^{-inx} \right) \\ &= \frac{\sinh(\pi)}{\pi} \left( 1 + \sum_{n=1}^{\infty} \frac{(-1)^n}{1+n^2} \left[ (1+in) e^{inx} + (1-in) e^{-inx} \right] \right) \\ &= \frac{\sinh(\pi)}{\pi} \left( 1 + \sum_{n=1}^{\infty} \frac{(-1)^n}{1+n^2} \left[ e^{inx} + e^{-inx} + in(e^{inx} - e^{-inx}) \right] \right) \\ &= \frac{\sinh(\pi)}{\pi} \left( 1 + 2 \sum_{n=1}^{\infty} \frac{(-1)^n}{1+n^2} \left[ \cos(nx) + i^2 n \sin(nx) \right] \right) \\ &= \frac{\sinh(\pi)}{\pi} \left( 1 + 2 \sum_{n=1}^{\infty} \frac{(-1)^n}{1+n^2} \left[ \cos(nx) - n \sin(nx) \right] \right). \end{aligned}$$

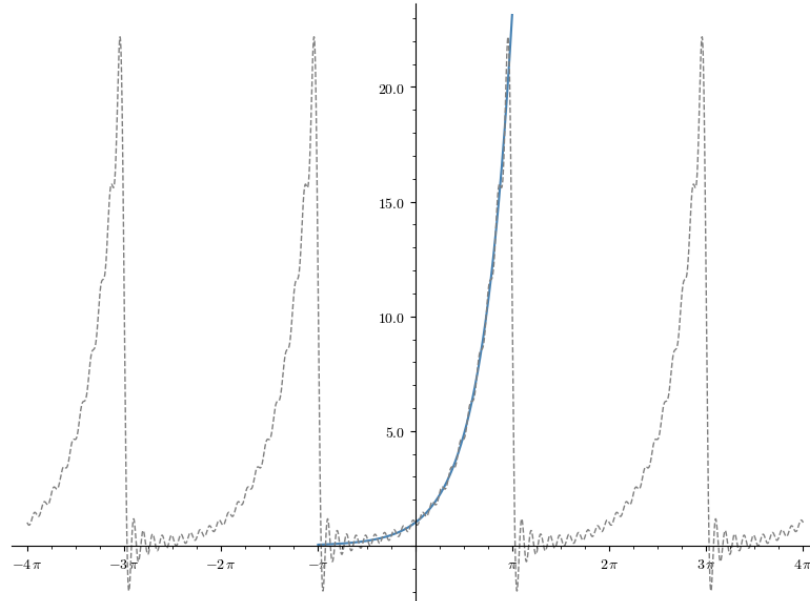
In order to plot  $f$  together with the 20th partial sum of the Fourier series, use the cell given here:

```
f=piecewise([[(-pi,pi), e**x]])
a=plot(f, x, -pi, pi, color="steelblue", thickness=1.5)
```

```

Fs20 = f.fourier_series_partial_sum(20,pi)
s20 = plot(Fs20, x, -4*pi, 4*pi,color="grey", linestyle="--")
(a+s20).show(ticks=pi, tick_formatter=pi, figsize=8)
    
```

Running this cell we obtain the following figure:



**7.A.18.** In this case the period is given by  $T = 2L = 1$ , i.e.,  $L = 1$ . Thus

$$F(x) = \sum_{n=-\infty}^{\infty} c_n e^{\frac{in\pi x}{L}} = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x}, \quad c_n = \frac{1}{2L} \int_{-L}^L e^{-x} e^{-\frac{in\pi x}{L}} dx = \frac{1}{2} \int_{-1}^1 e^{-x} e^{-in\pi x} dx.$$

We compute

$$\begin{aligned} c_n &= \frac{1}{2} \int_{-1}^1 e^{-(1+in\pi)x} dx = \frac{1}{2} \left[ \frac{e^{-(1+in\pi)x}}{-(1+in\pi)} \right]_{-1}^1 = -\frac{1}{2(1+in\pi)} \left( e^{-(1+in\pi)} - e^{(1+in\pi)} \right) \\ &= -\frac{1}{2(1+in\pi)} \left( e^{-1} e^{-in\pi} - e^1 e^{in\pi} \right) = \frac{(-1)^n}{(1+in\pi)} \frac{e^1 - e^{-1}}{2} = \frac{(-1)^n}{(1+in\pi)} \sinh(1). \end{aligned}$$

Here we used the identities  $e^{in\pi} = e^{-in\pi} = \cos(n\pi) = (-1)^n$ . Hence, the Fourier series of  $f$  has the form

$$F(x) = \sum_{n=-\infty}^{\infty} \left[ \frac{(-1)^n}{(1+in\pi)} \sinh(1) \right] e^{in\pi x} = \sinh(1) \sum_{n=-\infty}^{\infty} \frac{(-1)^n (1-in\pi)}{1+n^2\pi^2} e^{in\pi x}.$$

**7.A.24.** We need the even extension of  $f$  on  $(-\pi, \pi)$ , which reads by

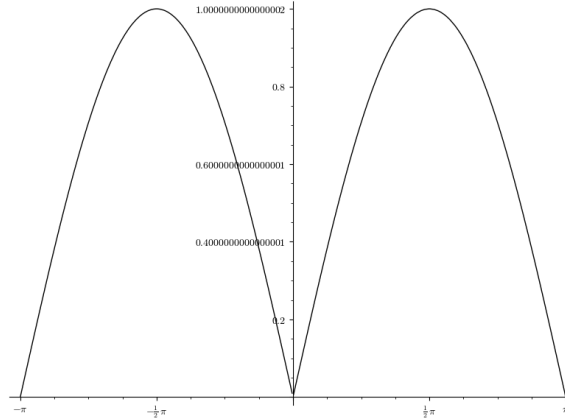
$$f_{\text{ev}}(x) = \begin{cases} f(x), & \text{if } 0 \leq x < \pi, \\ f(-x), & \text{if } -\pi < x < 0 \end{cases} = \begin{cases} \sin(x), & \text{if } 0 \leq x < \pi, \\ -\sin(x), & \text{if } -\pi < x < 0. \end{cases}$$

We may extend  $f_{\text{ev}}$  periodically all over  $\mathbb{R}$  by setting  $f_{\text{ev}}(x + 2\pi) = f_{\text{ev}}(x)$  for all the other  $x$ . Notice the values of this  $2\pi$ -periodic extension at  $\pm\pi$  coincide with  $f(\pi) = \sin(\pi) = 0$  (and similarly for  $\pm N\pi$ ,  $N \in \mathbb{N}$ ). Let us use Sage to illustrate this extension on  $(-\pi, \pi)$ :

```

f1(x)=sin(x); f2(x)=-sin(x)
f = piecewise([[(-pi, 0), f2], [RealSet.closed_open(0, pi), f1]])
p=f.plot(x, -pi, pi, exclude=[0], color="black")
p.show(figsize=8, ticks=pi/2, tick_formatter=pi)
    
```

Running this block we obtain the figure given here:



Now, for the cosine extension we have  $b_n = 0$  while the coefficients  $a_n$  are given by  $a_n = \frac{2}{\pi} \int_0^\pi \sin(x) \cos(nx) dx$ . Using the identity  $2 \sin(\theta) \cos(\varphi) = \sin(\theta + \varphi) + \sin(\theta - \varphi)$  we obtain

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_0^\pi \left( \sin((1+n)x) + \sin((1-n)x) \right) dx = \frac{1}{\pi} \left[ -\frac{\cos((1+n)x)}{n+1} - \frac{\cos((1-n)x)}{1-n} \right]_0^\pi \\ &= \frac{1}{\pi} \left( \frac{1 - \cos((n+1)\pi)}{n+1} + \frac{\cos((1-n)\pi) - 1}{n-1} \right) = \frac{1}{\pi} \left( \frac{1 - (-1)^{n+1}}{n+1} + \frac{(-1)^{n+1} - 1}{n-1} \right) \\ &= \frac{1}{\pi} \left( \frac{1 + (-1)^n}{n+1} - \frac{(-1)^n + 1}{n-1} \right) = \frac{(1 + (-1)^n)}{\pi} \left( \frac{1}{n+1} - \frac{1}{n-1} \right) = -\frac{2(1 + (-1)^n)}{\pi(n^2 - 1)}, \end{aligned}$$

which makes sense for all integers  $n > 1$ . To verify this expression in Sage one may add in the block presented above the following cell:

```
L=pi; n=var("n"); an=(2/L)*integral(sin(x)*cos(n*pi*x/L), x, 0, pi);
show("a_n=", an.expand())
```

Moreover, it is easy to see that

$$a_1 = \frac{2}{\pi} \int_0^\pi \sin(x) \cos(x) dx = 0, \quad a_0 = \frac{2}{\pi} \int_0^\pi \sin(x) dx = \frac{4}{\pi}.$$

All together, we have proved that

$$a_n = \begin{cases} \pi/4, & \text{if } n = 0, \\ 0, & \text{if } n \geq 1, n = 2k + 1 \text{ odd}, \\ -\frac{4}{(n^2-1)\pi}, & \text{if } n > 1, n = 2k \text{ even}. \end{cases}$$

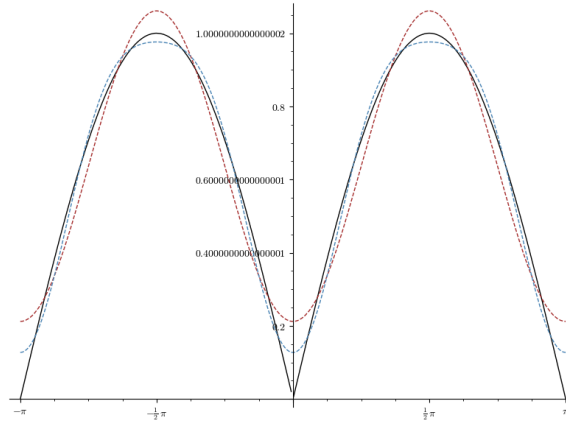
Therefore for all  $x \in (-\pi, \pi)$  by Dirichlet's condition we have that

$$\sin(x) = \frac{2}{\pi} - \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{\cos(2kx)}{4k^2 - 1} = \frac{2}{\pi} - \frac{4}{\pi} \left( \frac{\cos(2x)}{3} + \frac{\cos(4x)}{15} + \frac{\cos(6x)}{35} + \dots \right).$$

We may use Sage to illustrate the 3rd and 5th partial sums of the approximation, with brown and blue color respectively, together with the graph of  $f_{ev}$  for  $x \in (-\pi, \pi)$ . It suffices to add to the initial block the following one:

```
Fs3 = f.fourier_series_partial_sum(3, pi)
Fs5 = f.fourier_series_partial_sum(5, pi)
s3 = plot(Fs3, x, -pi, pi, color="brown", linestyle="--")
s5 = plot(Fs5, x, -pi, pi, linestyle="--", color="steelblue")
(p+s3+s5).show(ticks=pi/2, tick_formatter=pi, figsize=8)
```

Let us pose the figure that one obtains after executing the whole syntax together:



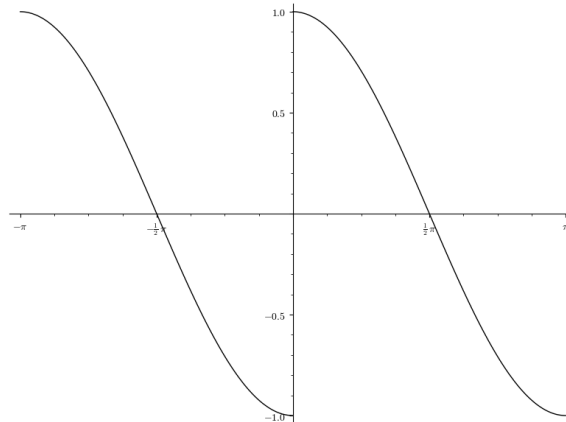
**7.A.25.** Here we need the odd extension of  $f$ , which is given by

$$f_{\text{odd}}(x) = \begin{cases} f(x), & \text{if } 0 < x < \pi, \\ -f(-x), & \text{if } -\pi < x < 0 \end{cases} = \begin{cases} \cos(x), & \text{if } 0 < x < \pi, \\ -\cos(x), & \text{if } -\pi < x < 0. \end{cases}$$

We may extend  $f_{\text{odd}}$  over the whole real line by setting  $f_{\text{odd}}(x + 2\pi) = f_{\text{odd}}(x)$  for  $x$  outside  $(-\pi, \pi)$ . At  $x = \pm\pi$  the values of this extension will coincide with the values  $f(\pm\pi) = \mp 1$ . Let us illustrate the extension for  $x \in (-\pi, \pi)$  by Sage, via the same method as above:

```
f1(x)=cos(x); f2(x)=-cos(x)
f = piecewise([(-pi, 0), f2], [(0, pi), f1])
p=f.plot(x, -pi, pi, exclude=[0])
p.show(figsize=8, ticks=pi/2, tick_formatter=pi)
```

Running this block we obtain the figure given here:



Now, since we are interesting in the sine expansion, we have necessarily  $a_n = 0$  for all  $n \in \mathbb{N}$ . Moreover, since  $L = \pi$  we see that  $b_n = \frac{2}{\pi} \int_0^\pi \cos(x) \sin(nx) dx$ . To compute this integral use the identity  $2 \sin(nx) \cos(x) = \sin((1+n)x) + \sin((n-1)x)$ , which gives

$$b_n = \frac{1}{\pi} \int_0^\pi (\sin((1+n)x) + \sin((n-1)x)) dx = -\frac{1}{\pi} \left[ \frac{\cos((n+1)x)}{n+1} + \frac{\cos((n-1)x)}{n-1} \right]_0^\pi = \frac{2n((-1)^n + 1)}{(n^2 - 1)\pi}.$$

This holds for all  $n > 1$  and can be verified in Sage by adding in the previous block the syntax given here:

```
L=pi; n=var("n")
bn=(2/L)*integral(cos(x)*sin(n*pi*x/L), x, 0, pi)
show("b_n=", bn.expand())
```

On the other hand, for  $n = 1$  we get

$$b_1 = \frac{2}{\pi} \int_0^\pi \cos(x) \sin(x) dx = \frac{1}{\pi} \int_0^\pi \sin(2x) dx = 0,$$

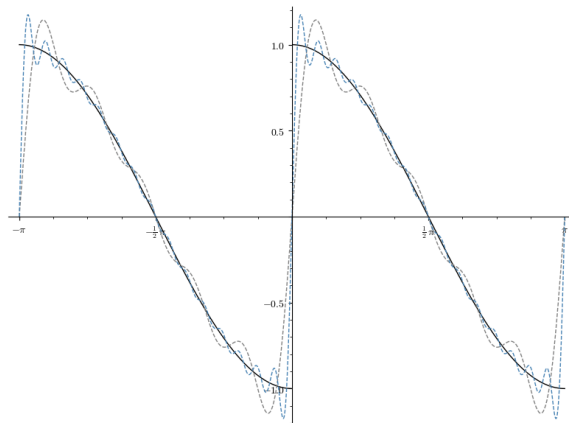
thus we have proved that

$$b_n = \begin{cases} 0, & \text{if } n = 1, \\ 0, & \text{if } n > 1, n = 2k + 1 \text{ odd}, \\ \frac{4n}{(n^2-1)\pi}, & \text{if } n > 1, n = 2k \text{ even}. \end{cases}$$

Therefore, one deduces that

$$\cos(x) = \sum_{k=1}^{\infty} \frac{8k}{(4k^2-1)\pi} \sin(2kx) = \frac{8}{3\pi} \sin(2x) + \frac{16}{15\pi} \sin(4x) + \frac{24}{35\pi} \sin(6x) + \dots$$

for all  $x \in (0, \pi)$ . Below are presented the approximations obtained for  $n = 10$  and  $n = 30$ , respectively, on the interval  $(-\pi, \pi)$ .



#### 7.B.4.

$$f_1 * f_2(t) = \begin{cases} t - \frac{t^2}{2} + 4 & \text{for } t \in [-2, -1] \\ 1 - t + \frac{1}{2} & \text{for } t \in [-1, 1] \\ \frac{t^2}{2} - 2t + 2 & \text{for } t \in [1, 2] \\ 0 & \text{otherwise.} \end{cases}$$

**7.B.14.** It is a good exercise on derivation and integration by parts. We may differentiate with respect to  $t$  inside the integral and  $\frac{d}{dt}f(t-x)$  can be interpreted as  $-\frac{d}{dx}f(t-x)$ .

**7.B.16.** Another good exercise on derivation and integration by parts. We may differentiate twice with respect to  $t$  inside the integral and  $f''(t-x)$  can be interpreted either as a derivative with respect to  $t$  or  $x$ .

**7.B.18.** The definition of  $\Gamma(t)$  reveals

$$\mathcal{L}(t^\alpha) = \int_0^\infty e^{-st} t^\alpha dt = \frac{1}{s^{\alpha+1}} \int_0^\infty e^{-x} x^\alpha dx = \frac{\Gamma(\alpha+1)}{s^{\alpha+1}}.$$

**7.B.19.** Integrate by parts to obtain

$$\mathcal{L}(g)(s) = \int_0^\infty t e^{-t} e^{-st} dt = \int_0^\infty t e^{-(s+1)t} dt = \lim_{t \rightarrow \infty} \left( \frac{t e^{-(s+1)t}}{-(s+1)} \right) - 0 - \int_0^\infty \frac{e^{-(s+1)t}}{-(s+1)} dt = - \left( \lim_{t \rightarrow \infty} \frac{e^{-(s+1)t}}{(s+1)^2} - \frac{e^0}{(s+1)^2} \right) = \frac{1}{(s+1)^2}.$$

Differentiating the Laplace transform of a general function  $-f$  (i. e., an improper integral) with respect to the parameter  $s$  gives

$$\left( \int_0^\infty -f(t) e^{-st} dt \right)' = \int_0^\infty -f(t) (e^{-st})' dt = \int_0^\infty t f(t) e^{-st} dt.$$

This means that the derivative of the Laplace transform  $\mathcal{L}(-f)(s)$  is the Laplace transform of the function  $tf(t)$ . The Laplace transform of the function  $y = \sinh t$  has already been determined as the function  $y = \frac{1}{s^2-1}$ . Therefore,

$$\mathcal{L}(h)(s) = \left( -\frac{1}{s^2-1} \right)' = \frac{2s}{(s^2-1)^2}.$$



We could also have determined  $\mathcal{L}(g)(s)$  this way.

**7.B.20.**

$$\begin{aligned} & \mathcal{L}(\cos \omega t)(s) + i\mathcal{L}(\sin \omega t)(s) = \mathcal{L}(e^{i\omega t})(s) \\ &= \int_0^\infty e^{i\omega t} e^{-st} dt = \int_0^\infty e^{(i\omega-s)t} dt \\ &= -\frac{1}{s-i\omega} \left[ e^{(i\omega-s)t} \right]_0^\infty \\ &= -\frac{1}{s-i\omega} \left( \lim_{t \rightarrow \infty} \frac{e^{i\omega t}}{e^{st}} - 1 \right) = \frac{1}{s-i\omega} = \frac{s+i\omega}{(s-i\omega)(s+i\omega)} \\ &= \frac{s}{s^2+\omega^2} + i\frac{\omega}{s^2+\omega^2} \end{aligned}$$

**7.C.23.** Recall the definition of the product metric of the Cartesian product  $X \times X$  where the distance is given as the maximum of the distance of the components. The claim follows directly from the triangle inequality and the topological definition of the continuity.

**7.D.5.** The orthogonal basis has the form  $\{x, -\frac{3}{\pi^2}x + \sin(x)\}$ , while the projection does not change the function  $\frac{1}{2}\sin(x)$  since it already lies in the space.

**7.D.6.** The orthogonal basis has the form  $\{\cos(x), \frac{4}{\pi}\cos(x) + x\}$ . Moreover, the required projection is given by  $3\pi/(\pi^4 - 24)(4\cos(x) + \pi x)$ . Notice that this is a very bad approximation.

**7.D.7.** We have already checked the orthogonality of the cosine terms in the solution to the example 7.A.4. The sine terms are obtained the same way since they are just shifts of cosines by  $\pi/2$  in the argument. The mixed couples provide an odd function to be integrated on a symmetric interval around the origin and so the integral also vanishes.

**7.D.10.** For  $n \neq 0$  we have

$$c_n = \frac{1}{2\pi} \int_0^\pi e^{-inx} dx = \frac{1}{2\pi} \left[ \frac{e^{-inx}}{-in} \right]_0^\pi = \frac{1}{2\pi} \frac{(1 - (-1)^n)}{in}.$$

Moreover,  $c_0 = \frac{1}{2\pi} \int_0^\pi dx = \frac{1}{2}$ . Thus

$$F(x) = \frac{1}{2} + \sum_{n=1}^\infty c_n e^{inx} + \sum_{n=-\infty}^{-1} c_n e^{inx} = \frac{1}{2} + \sum_{n=1}^\infty \frac{1}{2\pi} \frac{(1 - (-1)^n)}{in} e^{inx} + \sum_{n=-\infty}^{-1} \frac{1}{2\pi} \frac{(1 - (-1)^n)}{in} e^{inx}.$$

Next, in the second of the sums, replace  $n$  with  $-n$ . This gives

$$\begin{aligned} F(x) &= \frac{1}{2} + \frac{1}{2\pi} \sum_{n=1}^\infty \frac{(1 - (-1)^n)}{in} (e^{inx} - e^{-inx}) = \frac{1}{2} + \frac{1}{2\pi} \sum_{n=1}^\infty \frac{(1 - (-1)^n)}{in} 2i \sin(nx) \\ &= \frac{1}{2} + \frac{1}{\pi} \sum_{n=1}^\infty \frac{(1 - (-1)^n)}{n} \sin(nx). \end{aligned}$$

It is easy to see that when  $n$  is even, the terms are all 0. For  $n$  odd, put  $n = 2m - 1$  to obtain

$$F(x) = \frac{1}{2} + \frac{2}{\pi} \sum_{m=1}^\infty \left( \frac{1}{2m-1} \right) \sin(2m-1)x = \frac{1}{2} + \frac{2}{\pi} \left( \sin(x) + \frac{\sin(3x)}{3} + \frac{\sin(5x)}{5} + \dots \right).$$

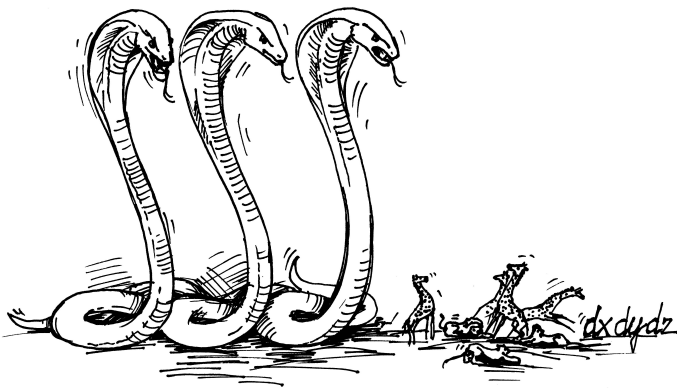
Notice that this expression is similar to the problem described in 7.1.9. In particular, we could use the result from 7.1.9 to derive our claim after an easy transformation.

**7.D.15.** Determine the convolution of the functions  $f_1$  and  $f_2$ , where

$$f_1 * f_2(t) = \begin{cases} \frac{(t+1)^2}{2} & \text{for } t \in [-1, 0] \\ \frac{1-t^2}{2} & \text{for } t \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

## Calculus with more variables

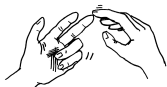
*one variable is not enough?  
– never mind, just recall vectors!*



### A. Multivariate functions

We start this chapter with a couple of easy examples to "grasp" a little multivariate functions.

**8.A.1.** Solve the system of inequalities. Mark the resulting area in the plane.



- a)  $x^2 + y^2 \leq 4$   
 $y \geq \frac{1}{x}$
- b)  $y \leq \arctan x$   
 $y \leq \frac{1}{x^2}$
- c)  $x^2 + (y - 1)^2 \geq 4$   
 $y + x^2 - 2x \geq 0$   
 $y \geq 0$

**Solution.** Whenever you have to solve an inequality of the form  $f(x, y) \geq 0$  ( $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  is a function of two variables, but the same method is valid for inequalities with more variables), you just consider the border curve  $f(x, y) = 0$ . This curve divides the plane into some areas. Then all points in any of the areas either satisfy the inequality or the whole area does not satisfy it. If we have a system of inequalities, we

At the beginning of our journey through the mathematical landscape, we saw that vectors can be manipulated nearly as easily as scalars. Now, we return to situations where the relations between objects are expressed with the help of more (yet still finitely many) parameters. This is really necessary when modeling processes or objects in practice, where functions  $\mathbb{R} \rightarrow \mathbb{R}$  of one variable are seldom adequate. At least, functions dependent on finitely many parameters are necessary, and the dependence of the change of the results on the parameters is often more important than the result itself. There is little need for brand new ideas. Many problems we encounter can be reduced to ones we can solve already. We return to the discussion of situations when the values of functions are described in terms of instantaneous changes. That is, we consider ordinary differential equations. In the next chapter, we consider partial differential equations and provide a gentle introduction to variational problems.

### 1. Functions and mappings on $\mathbb{R}^n$

**8.1.1. The world of functions.** In the sequel, the main objects are mappings between Euclidean spaces,  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ . We have seen many such examples already. The complex valued real functions correspond to  $n = 1, m = 2$ , while the power series converge inside of a circle in the complex plane, providing examples of  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ . We have also dealt with vector valued real functions, representing parametrized curves  $c : \mathbb{R} \rightarrow \mathbb{R}^n$  (see e.g. the paragraphs on curvatures and Frenet frames in 6.1.16 on page 526).

In linear algebra and geometry, we saw the linear and affine maps  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  defined with the help of matrices  $A \in \text{Mat}_{m,n}(\mathbb{R})$  and constant vectors  $y \in \mathbb{R}^m$ :

$$\mathbb{R}^n \ni x \mapsto y + Ax \in \mathbb{R}^m.$$

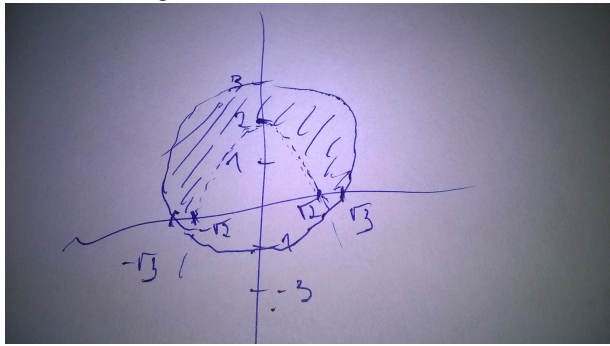
In coordinates, the value is given by the expression  $\sum_j a_{ij}x_j + y_i$ , where  $A = (a_{ij})$  and  $y = (y_i)$ .

Finally, the quadratic forms were mappings  $\mathbb{R}^n \rightarrow \mathbb{R}$  given by symmetric matrices  $S = (s_{ij})$  and the formula

$$\mathbb{R}^n \ni x \mapsto x^T S x \in \mathbb{R}.$$

In coordinates, the value is  $\sum_{i,j} s_{ij}x_i x_j$ .

solve each inequality separately and then intersect the result. In our cases we get



**8.A.2.** Determine the domain of the function  $\mathbb{R}^2 \rightarrow \mathbb{R}$ :

a)

$$\frac{xy}{y(x^3 + x^2 + x + 1)},$$

b)

$$\ln(x^2 - y^2),$$

c)

$$\ln(-x^2 - y^2),$$

d)

$$\arcsin(2\chi_{\mathbb{Q}}(x)),$$

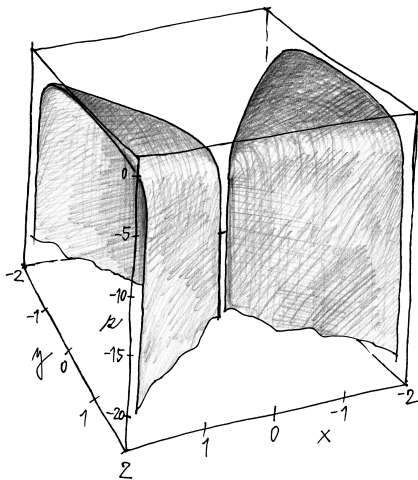
where  $\chi_{\mathbb{Q}}$  denotes the indicator function of the rational numbers,

e)

$$f(x, y, z) = \sqrt{\ln x \cdot \arcsin(y^2 z)}.$$

**Solution.** a) The formula correctly expresses a value iff the denominator of the fraction is non-zero. Therefore, the formula defines a function on the set  $\mathbb{R}^2 \setminus \{([x, 0], [-1, y]), x, y \in \mathbb{R}\}$ .

b) The formula is correct iff the argument of the logarithm is positive, i. e.,  $|x| > |y|$ . Therefore, the domain of this function is  $\{(x, y) \in \mathbb{R}^2, |x| > |y|\}$ . You can see the graph of this function in the picture.



c) This formula is again a composition of a logarithm and a polynomial of two variables. However, the polynomial  $-x^2 -$

In general, all such mappings  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are composed of  $m$  components of functions  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ . So we start with this case.

**8.1.2. Multivariate functions.** We can stress the dependence on the variables  $x_1, \dots, x_n$  by writing the functions as

$$f(x_1, x_2, \dots, x_n) : \mathbb{R}^n \rightarrow \mathbb{R}.$$

The goal is to extend methods for monitoring the values of functions and their changes for this situation.

We speak about *functions of more variables* or, more compactly, *multivariate functions*.

We often work with the cases  $n = 2$  or  $n = 3$  so that the concepts being introduced are easier to understand. In these cases, letters like  $x, y, z$  are used instead of numbered variables. This means that a function  $f$  defined in the “plane”  $\mathbb{R}^2$  is denoted

$$\mathbb{R}^2 \ni (x, y) \mapsto f(x, y) \in \mathbb{R},$$

and, similarly, in the “space”  $\mathbb{R}^3$

$$\mathbb{R}^3 \ni (x, y, z) \mapsto f(x, y, z) \in \mathbb{R}.$$

Just as in the case of univariate functions, the domain  $A \subset \mathbb{R}^n$  on which the function in question is defined needs to be considered. When examining a function given by a concrete formula, the first task is often to find the largest domain on which the formula makes sense.

It is also useful to consider the *graph* of a multivariate function, i. e., the subset  $G_f \subset \mathbb{R}^n \times \mathbb{R} = \mathbb{R}^{n+1}$ , defined by

$$G_f = \{(x_1, \dots, x_n, f(x_1, \dots, x_n)); (x_1, \dots, x_n) \in A\},$$

where  $A$  is the domain of  $f$ . For instance, the graph of the function defined in the plane by the formula

$$f(x, y) = \frac{x + y}{x^2 + y^2}$$

is quite a smooth surface, see the illustration below. The maximal domain of this function consists of all the points of the plane except for the origin  $(0, 0)$ .

When defining the function, and especially when drawing its graph, fixed *coordinates* are used in the plane. Fixing the value of either of the coordinates, implies only one variable remains. Fixing the value of  $x$ , for example, gives the mapping

$$\mathbb{R} \rightarrow \mathbb{R}^3, y \mapsto (x, y, f(x, y)),$$

i.e., a *curve* in the space  $\mathbb{R}^3$ . Curves are vector functions of one variable, already worked with in chapter six (see 6.1.14). The images of the curves for some fixed values of the coordinates  $x$  and  $y$  are depicted by lines in the illustration.

add the “coordinate lines” in the picture

$y^2$  takes on only non-positive real values, where the logarithm is undefined (as a function  $\mathbb{R} \rightarrow \mathbb{R}$ ).

d) This formula correctly defines a value iff the argument of the arc sine lies in the interval  $[-1, 1]$ , which is broken by exactly those pairs  $(= [x, y] \in \mathbb{R}^2$  whose first component is rational. The formula thus defines a function on the set  $\{[x, y], x \in \mathbb{R} \setminus \mathbb{Q}\}$ .

e) The argument of the square root must be non-negative, that is either the image of the logarithm is positive and the image of arcsine as well, or both images are negative. Thus we get that the domain is the set

$$\{[x, y, z] \in \mathbb{R}^3; (x \geq 1 \wedge y \neq 0 \wedge 0 \leq z \leq \frac{1}{y^2}) \vee (x \in (0, 1) \wedge y \neq 0 \wedge -\frac{1}{y^2} \leq z \leq 0) \vee (x > 0 \wedge y = 0)\}.$$

□

In the following examples  $k([x, y]; r)$  means a circle with the center  $[x, y]$  and the radius  $r$ .

**8.A.3.** Determine the domain of the function  $f$  and mark the resulting area in the plane:

- i)  $f(x, y) = \sqrt{(x^2 + y^2 - 1)(4 - x^2 - y^2)}$ ,
- ii)  $f(x, y) = \sqrt{1 - x^2} + \sqrt{1 - y^2}$ ,
- iii)  $f(x, y) = \sqrt{\frac{x^2 + y^2 - x}{2x - x^2 - y^2}}$ ,
- iv)  $f(x, y) = \arcsin \frac{x}{y} - \frac{1}{|y| - |x|}$ ,
- v)  $f(x, y) = \sqrt{1 - x^2 - 4y^2}$ ,
- vi)  $f(x, y, z) = \sqrt{1 - \frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2}}$ .

**Solution.** a) It has to hold  $(x^2 + y^2 - 1 \geq 0, 4 - x^2 - y^2 \geq 0)$  or  $(x^2 + y^2 - 1 \leq 0, 4 - x^2 - y^2 \leq 0)$ , that is  $(x^2 + y^2 \geq 1, x^2 + y^2 \leq 4)$  or  $(x^2 + y^2 \leq 1, x^2 + y^2 \geq 4)$ , which is an annulus between the circles  $k([0, 0]; 1)$  and  $k([0, 0]; 2)$ .

b) It is a circle with the center  $[0, 0]$  and vertices  $[\pm 1, \pm 1]$

c) The area between circles  $k([\frac{1}{2}, 0]; \frac{1}{2})$  and  $k([1, 0]; 1)$ , the smaller circle belongs to the area, the bigger one does not.

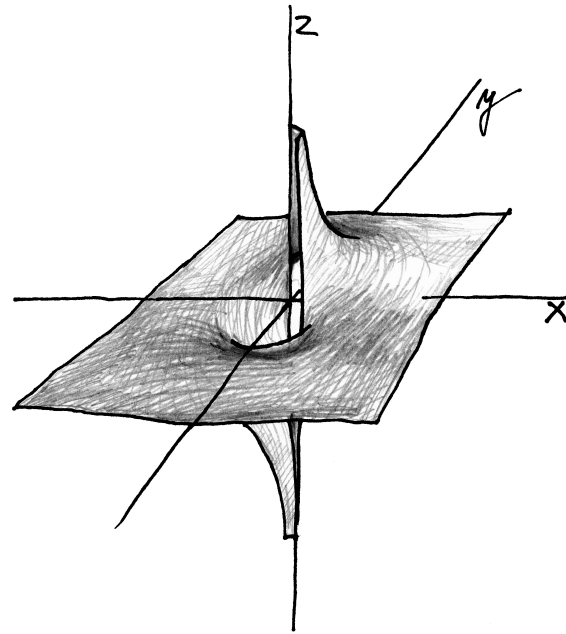
d) The area between the lines  $y = x$  and  $y = -x$  (without these lines).

e) The ellipse (together with the inner space) with the center  $[0, 0]$ , with the major axis lying on the  $x$ -axis with the major radius  $a = 1$ , and the minor axis on the  $y$ -axis with the minor radius  $b = \frac{1}{2}$ .

f) The ellipsoid (with the inner space) with the center  $[0, 0, 0]$  a semiaxes lying on the  $x, y, z$  axis respectively, with radii  $a, b$ , and  $c$ . □

### B. The topology of $E_n$

In the previous chapter, we have defined general metric spaces and we have studied especially metric spaces consisting of the set of functions. As we have already seen in the previous chapter, many metrics can be defined on the space  $\mathbb{R}^n$  (or on its subsets). For instance, considering a map of a



### 8.1.3. Euclidean spaces.



In the case of functions of one variable, the entire differential and integral calculus is based on the concepts of convergence, open neighbourhoods, continuity, and so on. In the last part of chapter seven, these concepts were generalized for the metric spaces, rather than only for the Euclidean spaces  $\mathbb{R}^n$ . Before proceeding it is appropriate to revise these ideas, and do further reading if necessary. We present a brief summary:

The Euclidean space  $E_n$  is perceived as a set of points in  $\mathbb{R}^n$  without any choice of coordinates, and its modelling vector space  $\mathbb{R}^n$  is considered to be the vector space of all increments that can be added to the points of the space  $E_n$  (the modelling vector space).

Moreover, the standard scalar product

$$u \cdot v = \sum_{i=1}^n x_i y_i,$$

is selected on  $\mathbb{R}^n$ , where  $u = (x_1, \dots, x_n)$  and  $v = (y_1, \dots, y_n)$  are arbitrary vectors. This gives a *metric* on  $E_n$ , i.e. a function describing the distance  $\|P - Q\|$  between pairs of points  $P, Q$  by the formula

$$\|P - Q\|^2 = \|u\|^2 = \sum_{i=1}^n x_i^2,$$

where  $u$  is the vector which yields the point  $P$  when added to the point  $Q$ . In the plane  $E_2$ , for instance, the distance between the points  $P_1 = (x_1, y_1)$  and  $P_2 = (x_2, y_2)$  is given by

$$\|P_1 - P_2\|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2.$$

The metric defined in this manner satisfies the triangle inequality for every triple of points  $P, Q, R$ :

$$\|P - R\| = \|(P - Q) + (Q - R)\| \leq \|P - Q\| + \|Q - R\|.$$

state as a subset of  $\mathbb{R}^2$ , the distance of two points may be defined as the time necessary to get from one of the points to the other by public transport or on foot. In France, for example, the shortest paths between most pairs of points in this metric are far from line segments. In this chapter we will focus on the space  $E_n$ , that is  $\mathbb{R}^n$  with the usual metric (distance) known to the mankind for a long time. The property, that the shortest path between any two points of this space is the line segment connecting them could be seen as the defining property (for example the above example does not satisfy it). Let us examine the space  $E_n$  in more detail.

**8.B.1.** Show that every non-empty proper subset of  $E_n$  has a boundary point (which need not lie in it).

**Solution.** Let  $U \subset E_n$  be a non-empty subset with no boundary point. Consider a point  $X \in U$ , a point  $Y \in U' := E_n \setminus U$ , and the line segment  $XY \subset E_n$ . Intuitively, going from  $X$  to  $Y$  along this segment, we must once get from  $U$  to  $U'$ , and this can happen only at a boundary point (everyone who has ever been to a foreign country is surely well acquainted with this fact). Formally, let  $A$  be the point of  $XY$  for which  $|XA| = \sup\{|XZ|, XZ \in U\}$  (apparently, there is exactly one such point on the segment  $XY$ ). This point is a boundary point of  $U$ : it follows from the definition of  $A$  that any line segment  $XB$  (with  $B \in XA$ ) is contained in  $U$ ; in particular,  $B \in U$ . However, if there were a neighborhood of  $A$  contained in  $U$ , then there would exist a part of the line segment  $XY$  longer than  $XA$  which would be contained in  $U$ , which contradicts the definition of the point  $A$ . Therefore, any neighborhood of the point  $A$  contains a point from  $U$  as well as a point from  $E_n \setminus U$ .  $\square$

**8.B.2.** Prove that the only non-empty clopen (both closed and open) subset of  $E_n$  is  $E_n$  itself.

**Solution.** It follows from the above exercise 8.B.1 that every non-empty proper subset  $U$  of  $E_n$  has a boundary point. If  $U$  is closed, then it is equal to its closure; therefore, it contains all of its boundary points. However, an open set (by definition) cannot contain a boundary point.  $\square$

**8.B.3.** Show that the space  $E_n$  cannot be written as the union of (at least two) disjoint non-empty open sets.

**Solution.** Suppose that  $E_n$  can be expressed thus, i. e.,  $E_n = \cup_{i \in I} U_i$ , where  $I$  is an index set. Let us fix a set  $U$  from this union. Then, we can write  $E_n = U \cup \bar{U}$ , where both  $U$  and  $\bar{U}$  (being a union of open sets) are open. However, they are also complements of open sets; therefore, they are closed as well. This contradicts the result of the previous exercise 8.B.2.  $\square$

**8.B.4.** Prove or disprove: a union of (even infinitely many) closed subsets of  $E^n$  is a closed subset of  $E^n$ .

**Solution.** The proposition does not hold. As a counterexample, consider the union

$$\bigcup_{i=3}^{\infty} \left[ \frac{1}{i}, 1 - \frac{1}{i} \right]$$

See 3.4.3(1) in geometry, or the axioms of metrics in 7.3.1, or the same inequality 5.2.2(2) for complex scalars. The concepts defined for real and complex scalars and discussed for metric spaces in detail can be carried over (extended) with no problem for the points  $P_i$  of any Euclidean space:

TOPOLOGY AND METRIC IN EUCLIDEAN SPACES

- (1) *a Cauchy sequence*: a sequence of points  $P_i$  such that for every fixed  $\varepsilon > 0$ ,  $\|P_i - P_j\| < \varepsilon$  holds for all indices but for finitely many exceptional points  $P_k$ ;
- (2) *a convergent sequence*: a sequence of points  $P_i$  converges to a point  $P$  if and only if for every fixed  $\varepsilon > 0$ ,  $\|P_i - P\| < \varepsilon$  holds for all but finitely many indices  $i$ ; the point  $P$  is then called the *limit* of the sequence  $P_i$ ;
- (3) *a limit point  $P$  of a set  $A \subset E_n$* : there exists a sequence of points in  $A$  converging to  $P$  and different from  $P$ ;
- (4) *a closed set*: contains all of its limit points;
- (5) *an open set*: its complement is closed;
- (6) *an open  $\delta$ -neighbourhood of a point  $P$* : the set  $\mathcal{O}_\delta(P) = \{Q \in E_n; \|P - Q\| < \delta\}$ ,  $\delta \in \mathbb{R}$ ,  $\delta > 0$ ;
- (7) *a boundary point  $P$  of a set  $A$* : every  $\delta$ -neighbourhood of  $P$  has non-empty intersection with both  $A$  and the complement  $E_n \setminus A$ ;
- (8) *an interior point  $P$  of a set  $A$* : there exists a  $\delta$ -neighbourhood of  $P$  which lies inside  $A$ ;
- (9) *a bounded set*: lies inside some  $\delta$ -neighbourhood of one of its points (for a sufficiently large  $\delta$ );
- (10) *a compact set*: both closed and bounded.
- (11) *limit of a mapping*:  $a \in \mathbb{R}^m$  is the limit of function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  in a limit point  $x_0$  of its domain  $A$ , if for each  $\varepsilon > 0$ , there is a  $\delta$ -neighbourhood  $U$  of  $x_0$ , such that  $\|f(x) - a\| < \varepsilon$  for all  $x \in U$ ; this happens if and only if for each sequence  $x_n \in A$  converging to  $x_0$ , the values  $f(x_n)$  converge to  $a$ .
- (12) *continuity*: mapping  $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$  is continuous in  $x_0 \in A$  if the limit  $\lim_{x \rightarrow x_0} f(x)$  exists and equals to  $f(x_0)$ ; the mapping  $f$  is continuous on  $A$ , if it is continuous in all points in  $A$ .

Both the first and second items deal with norms of differences of points approaching zero. Since the square of the norm is the sum of squares of the individual components, it is clear that this happens if and only if the individual components approach zero. In particular, the sequences of points  $P_i$  are Cauchy or convergent if and only if these properties are possessed by the real sequences obtained from the particular coordinates of the points  $P_i$  in every Cartesian coordinate system. Therefore, it also follows from Lemma 5.2.3 that every Cauchy sequence of points in  $E_n$  is convergent. Especially,  $E_n$  is a complete metric space.

Similarly, the mappings from the item (11) are  $m$ -tuples of the component functions and the limits are given as the  $m$ -tuples of limits of these components.

Recall some further results already discussed at the more general level of the metric spaces in chapter seven:

hodil by se obrazek na ilustraci pojmu, napr. obr. 1

of closed subsets of  $\mathbb{R}$ , which is equal to the open interval  $(0, 1)$ .  $\square$

**8.B.5.** Prove or disprove: an intersection of (even infinitely many) open subsets of  $E^n$  is an open subset of  $E^n$ .

**Solution.** The proposition does not hold in general. As a counterexample, consider the intersection

$$\bigcap_{i=2}^{\infty} \left(1 - \frac{1}{i}, 1 + \frac{1}{i}\right)$$

of open subsets of  $\mathbb{R}$ , which is equal to the closed singleton  $\{1\}$ .  $\square$

**8.B.6.** Consider the graph of a continuous function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  as a subset of  $E_3$ . Determine whether this subset is open, closed, and compact, respectively.

**Solution.** The subset is not open since any neighborhood of a point  $[x_0, y_0, f(x_0, y_0)]$  contains a segment of the line  $x = x_0, y = y_0$ . However, there is a unique point of the graph of the function on this segment, and that is the point  $[x_0, y_0, f(x_0, y_0)]$ .

The continuity of  $f$  implies that the subset is closed – we will show that every convergent sequence of points of the graph of  $f$  converges to a point which also lies in the graph: If such a sequence is convergent in  $E_3$ , then it must converge in every component, so the sequence  $\{[x_n, y_n]\}_{n=1}^{\infty}$  is convergent in  $\mathbb{R}^2$ . Let us denote this limit by  $[a, b]$ . Then, it follows from the definition of continuity that its function values at the points  $[x_n, y_n]$  must converge to the value  $f(a, b)$ . However, this means that the sequence  $\{[x_n, y_n, f(x_n, y_n)]\}_{n=1}^{\infty}$  converges to the point  $[a, b, f(a, b)]$ , which belongs to the graph of the function  $f$ . Therefore, the graph is a closed set.

The subset is closed, yet it is not compact since it is not bounded (its orthogonal projection onto the coordinate plane  $xy$  is the whole  $\mathbb{R}^2$ ). (A subset of  $E_n$  is compact iff it is both closed and bounded.)  $\square$

And now let us study the limits of functions (a limit is defined thanks to the topology of  $E_n$ , see 8.1.3)

### C. Limits and continuity of multivariate functions

If we approach limits of multivariate functions, there is one fact we have to deal with:

Let us emphasize that there is no analogy of L'Hospital rule for multivariate functions. Counting limits  $\frac{0}{0}$  or  $\frac{\infty}{\infty}$ , we have to be "clever".

In one dimension we can approach a point either from right or left (and the limit in the point exists, if both one-sided limits exist and are equal to each other). In more dimensions we can approach a point from infinitely many directions, and the limit in the point exists, iff limits of the function narrowed to any path leading to the point exist and must be equal to each other.

The easiest way to obtain a limit is (as with the functions of one variable) to plug in the given point to the function prescription and if we get a meaningful expression, we are done.

A mapping is continuous if and only if its preimages of open sets are open (check this carefully!). Further, each continuous function on a compact set  $A$  is uniformly continuous, bounded and attains its maximum and minimum, cf. the paragraph 7.3.14 on the page 682.

The reader should make an appropriate effort to read the paragraphs 3.4.3, 5.2.5–5.2.8, 7.3.3–7.3.5, and 7.3.12 as well as try to think out/recall the definitions and connections of all these concepts.



**8.1.4. Compact sets.** Working with general open, closed, or compact sets could seem useless in the case of the real line  $E_1$  since intervals are almost always used.

In the case of metric spaces in the last part of chapter seven, the ideas are complicated at first sight. However, the same approach is easy in the case of Euclidean spaces  $\mathbb{R}^n$ . It is also very useful and important (and it is, of course, a special case of general metric spaces).

Just as in the case of  $E_1$  or  $E_2$ , we deal with the open covers of sets (i.e., systems of open sets containing the given sets), and Theorem 5.2.8 is again true (with mere reformulations):

**Theorem.** Subsets  $A \subset E_n$  of Euclidean spaces satisfy:

- (1)  $A$  is open if and only if it is a union of a countable (or finite) system of  $\delta$ -neighbourhoods,
- (2) every point  $a \in A$  is either interior or boundary,
- (3) every boundary point of  $A$  is either an isolated or a limit point of  $A$ ,
- (4)  $A$  is compact if and only if every infinite sequence contained in it has a subsequence converging to a point in  $A$ ,
- (5)  $A$  is compact if and only if each of its open covers contains a finite subcover.

**PROOF.** The proof from 5.2.8 can be reused without changes in the case of claims (1)–(3), yet now the concepts have to be perceived in a different way, and the "open intervals" are substituted with multidimensional  $\delta$ -neighbourhoods of appropriate points.



However, the proof of the fourth and fifth claims has to be adjusted properly. Actually, we proved the claims there for  $\mathbb{R}$  and  $\mathbb{C}$ , thus in dimensions one and two. Thus the reader may either extend the two-dimensional reasoning, or to rewrite the proof of the corresponding propositions for general metric spaces in 7.3.12, while noticing the parts which can be simplified for Euclidean spaces.  $\square$

**8.1.5. Curves in  $E_n$ .** Almost all the discussion about limits, derivatives, and integrals of functions in chapters 5 and 6 concerned functions of a real variable and real or complex values since only the triangle inequality valid for the magnitudes of the real and complex numbers is used. This argument can be carried over to any function of a real variable with values in a Euclidean



Otherwise we can get "undeterminate" expression. There are some tricks, we can use to count such a limit:

- (1) factorize the numerator or the denominator according to some known formula and then reduce,
- (2) expand the numerator and the denominator with an appropriate term and then shorten,
- (3)  $\frac{\text{bounded expression}}{\infty} = 0, 0 \cdot (\text{bounded expression}) = 0;$
- (4) use an appropriate substitution to get a limit of a function of one variable
- (5) try polar coordinates

$$\begin{aligned} x &= r \cos \varphi, \\ y &= r \sin \varphi \end{aligned}$$

(it usually works with the expression  $x^2 + y^2$ , we have  $x^2 + y^2 = r^2 \cos^2 \varphi + r^2 \sin^2 \varphi = r^2(\cos^2 \varphi + \sin^2 \varphi) = r^2$ , which is independent of  $\varphi$ );

- (6) try  $y = kx$  or  $y = kx^2$  or generally  $x = f(k)$  a  $y = g(k)$  (to prove the non-existence of the limit: if the limit after the substitution depends on  $k$ , the original limit does not exist)

8.C.1.  $\lim_{(x,y) \rightarrow (e^2,1)} \frac{\ln x}{y}$

8.C.2.  $\lim_{(x,y) \rightarrow (4,4)} \frac{\sqrt{x}-\sqrt{y}}{x-y}$

8.C.3.  $\lim_{(x,y) \rightarrow (1,\infty)} \frac{\cos y}{x+y}$

8.C.4.  $\lim_{(x,y) \rightarrow (0,2)} \frac{e^{xy}-1}{x}$

8.C.5.  $\lim_{(x,y) \rightarrow (\infty,\infty)} \frac{x^2+y^2}{x^4+y^4}$

8.C.6.  $\lim_{(x,y) \rightarrow (0,0)} \frac{x^2+y^2}{x+y}$

8.C.7.  $\lim_{(x,y) \rightarrow (0,0)} \frac{x^2-y^2}{x^2+y^2}$

8.C.8.  $\lim_{(x,y) \rightarrow (\infty,\infty)} \left(\frac{2xy}{x^2+y^2}\right)^{x^2}$

8.C.9.  $\lim_{(x,y) \rightarrow (1,1)} \frac{x+y}{\sqrt{x^2+y^2}}$

8.C.10.  $\lim_{(x,y) \rightarrow (0,0)} \frac{x^2+y^2}{\sqrt{x^2+y^2+1}-1}$

8.C.11.  $\lim_{(x,y) \rightarrow (0,0)} xy^2 \cos \frac{1}{xy^2}$

8.C.12.  $\lim_{(x,y) \rightarrow (0,0)} \frac{\sin xy}{x}$

8.C.13.  $\lim_{(x,y) \rightarrow (0,0)} \frac{x^3+y^3}{x^2+y^2}$

8.C.14.  $\lim_{(x,y) \rightarrow (\infty,\infty)} (x^2 + y^2)e^{-(x+y)}$

8.C.15.  $\lim_{(x,y) \rightarrow (\infty,1)} \left(1 + \frac{1}{x}\right)^{\frac{x^2}{x+y}}$

8.C.16.  $\lim_{(x,y) \rightarrow (0,0)} \frac{xy}{x^2+y^2}$

8.C.17.  $\lim_{(x,y) \rightarrow (0,0)} \frac{1-\cos(x^2+y^2)}{(x^2+y^2)xy}$

8.C.18. Prove that  $\lim_{(x,y) \rightarrow (0,0)} \frac{-y}{x^2-y}$  does not exist.

8.C.19. Prove that

$$\lim_{(x,y) \rightarrow (1,-2)} \frac{2x + xy - y - 2}{x^2 + y^2 - 2x + 4y + 5}$$

does not exist.

space  $\mathbb{R}^n$ . Several tools for the work with curves are introduced in paragraphs 6.1.14–6.1.12.

For every (parametrized) curve<sup>1</sup>, that is, a mapping  $c = (c_1(t), \dots, c_n(t)) : \mathbb{R} \rightarrow \mathbb{R}^n$  in an  $n$ -dimensional space, the concepts simply extend the ideas from the univariate functions with some extra thoughts:

First note that both the limit and the derivative of curves make sense in an affine space even without selecting the coordinates (where the limit is again a point in the original space, while the derivative is a vector in the modeling vector space!).

In the case of an integral, curves are considered in the vector space  $\mathbb{R}^n$ . The reason for this can be seen even in the case of one dimension, where the origin is needed to be able to see the “area under the graph of a function”.

It is apparent that limits, derivatives, and integrals have to be considered via the  $n$  individual coordinate components in  $\mathbb{R}^n$ . In particular, their existence is determined in the same way:

BASIC CONCEPTS FOR CURVES

(1) *limit*:

$$\lim_{t \rightarrow t_0} c(t) = \left( \lim_{t \rightarrow t_0} c_1(t), \dots, \lim_{t \rightarrow t_0} c_n(t) \right) \in \mathbb{R}^n$$

(2) *derivative*:

$$\begin{aligned} c'(t_0) &= \lim_{t \rightarrow t_0} \frac{1}{|t - t_0|} \cdot (c(t) - c(t_0)) \\ &= (c'_1(t_0), \dots, c'_n(t_0)) \in \mathbb{R}^n \end{aligned}$$

(3) *integral*:

$$\int_a^b c(t) dt = \left( \int_a^b c_1(t) dt, \dots, \int_a^b c_n(t) dt \right) \in \mathbb{R}^n.$$

We can directly formulate the analogy of the connection between the Riemann integral and the antiderivative for curves in  $\mathbb{R}^n$  (see 6.2.9):

**Proposition.** *Let  $c$  be a curve in  $\mathbb{R}^n$ , continuous on an interval  $[a, b]$ . Then its Riemann integral  $\int_a^b c(t) dt$  exists. Moreover, the curve*

$$C(t) = \int_a^t c(s) ds \in \mathbb{R}^n$$

*is well-defined, differentiable, and  $C'(t) = c(t)$  for all values  $t \in [a, b]$ .*

It is not simple to extend the mean value theorem and, in general, the Taylor’s expansion with remainder, see 5.3.9 and 6.1.3. They can be applied in a selected coordinate system to the particular coordinate functions of a differentiable function  $c(t) = (c_1(t), \dots, c_n(t))$  on a finite interval  $[a, b]$ . In the case of the mean value theorem, for instance, there are numbers  $t_i$  such that

$$c_i(b) - c_i(a) = (b - a) \cdot c'_i(t_i), \quad i = 1, \dots, n.$$

<sup>1</sup>In geometry, one often makes a distinction between a curve as a subset of  $E_n$  and its parametrization  $\mathbb{R} \rightarrow \mathbb{R}^n$ . The word “curve”, means exclusively the parametrized curve here.

**Solution.** The lines through  $[1, -2]$  have the equation  $y = kx - k - 2$ . As we approach  $[1, -2]$  along one of these lines, we get the limit  $\frac{k}{1+k^2}$ , which is different for different  $k$ , thus the limit does not exist.  $\square$

Let us recall, that a function is continuous in points, where the limit exists and is equal to the function value.

**8.C.20.** Find the discontinuity points of  $f(x, y) = \frac{2x-5y}{x^2+y^2-1}$ .  $\circ$

**8.C.21.** Find the discontinuity points of  $f(x, y) = \frac{\sin(x^2y+xy^2)}{\cos(x-y)}$ .  $\circ$

**8.C.22.** Find the discontinuity points of

$$f(x, y) = \begin{cases} \frac{x^3+y^3}{x^2+y^2} & \text{pro } [x, y] \neq [0, 0], \\ 0 & \text{pro } [x, y] = [0, 0]. \end{cases}$$

$\circ$

**D. Tangent lines, tangent planes, graphs of multivariate functions**

**8.D.1.** A car is moving at velocity given by the vector  $(0, 1, 1)$ . At the initial time  $t = 0$ , it is situated at the point  $[1, 0, 0]$ . The acceleration of the car in time  $t$  is given by the vector  $(-\cos t, -\sin t, 0)$ . Describe the dependency of the position of the car upon the time  $t$ .

**Solution.** As we have already discussed in paragraph 8.1.5, we got acquainted with the means of solving this type of problem as early as in chapter 6. Notice that the “integral curve”  $C(t)$  from the theorem of paragraph 8.1.5 starts at the point  $(0, 0, 0)$  (in other words,  $C(0) = (0, 0, 0)$ ). In the affine space  $\mathbb{R}^n$ , we can move it so that it starts at an arbitrary point, and this does not change its derivative (this is performed by adding a constant to every component in the parametric equation of the curve). Therefore, up to the movement, this integral curve is determined uniquely (nothing else than constants can be added to the components without changing the derivative). When we integrate the curve of acceleration, we get the curve of velocity  $(-\sin t, \cos t - 1, 0)$ . Considering the initial velocity as well, we obtain the velocity curve of the car:  $(-\sin t, \cos t, 1)$  (we shifted the curve of the vector  $(0, 1, 1)$ , i. e., so that now the velocity curve at time  $t = 0$  agrees with the given initial velocity). Further integration leads to the curve  $(\cos t - 1, \sin t, t)$ . Shifting this of the vector  $(1, 0, 0)$  then fits with the initial position of the car. Therefore, the car moves along the curve  $[\cos t, \sin t, t]$  (this curve is called a helix).  $\square$

**8.D.2.** Determine both the parametric and implicit equations of the tangent line to the curve  $c : \mathbb{R} \rightarrow \mathbb{R}^3$ ,  $c(t) = (c_1(t), c_2(t), c_3(t)) = (t, t^2, t^3)$  at the point which corresponds to the parameter’s value  $t = 1$ .

**Solution.** The value  $t = 1$  corresponds to the point  $c(1) = [1, 1, 1]$ . The derivatives of the particular components are  $c'_1(t) = 1$ ,  $c'_2(t) = 2t$ ,  $c_3(t) = 3t^2$ . The values of

These numbers  $t_i$  are distinct in general, so the difference vector of the boundary points  $c(b) - c(a)$  cannot be expressed as a multiple of the derivative of the curve at a single point, in general.

For example, for a differentiable curve  $c(t)$  in the plane  $E_2$ ,  $c(t) = (x(t), y(t))$

$$\begin{aligned} c(b) - c(a) &= (x'(\xi)(b-a), y'(\eta)(b-a)) \\ &= (b-a) \cdot (x'(\xi), y'(\eta)) \end{aligned}$$

for two (in general different) values  $\xi, \eta \in [a, b]$ . However, this reasoning is still sufficient for the following estimate:

**Lemma.** If  $c$  is a curve in  $E_n$  with continuous derivative on a compact interval  $[a, b]$ , then for all  $a \leq s \leq t \leq b$

$$\|c(t) - c(s)\| \leq \sqrt{n}(\max_{r \in [a,b]} \|c'(r)\|) |t - s|.$$

**PROOF.** Direct application of the mean value theorem gives for appropriate points  $r_i$  inside the interval  $[s, t]$  the following:

$$\begin{aligned} \|c(t) - c(s)\|^2 &= \sum_{i=1}^n (c_i(t) - c_i(s))^2 \leq \sum_{i=1}^n (c'_i(r_i)(t-s))^2 \\ &\leq (t-s)^2 \sum_{i=1}^n \max_{r \in [s,t]} c'_i(r)^2 \\ &\leq n(\max_{r \in [s,t], i=1, \dots, n} |c'_i(r)|)^2 (t-s)^2 \\ &\leq n \max_{r \in [s,t]} \|c'(r)\|^2 (t-s)^2. \end{aligned}$$

$\square$

Another important concept is the *tangent vector* to a curve  $c : \mathbb{R} \rightarrow E_n$  at a point  $c(t_0) \in E_n$ . It is defined as the vector in the modelling vector space  $\mathbb{R}^n$  given by the derivative  $c'(t_0) \in \mathbb{R}^n$ .

Consider  $c$  to be the path of an object moving in the space in time. Then the tangent vector at a point  $t_0$  can be perceived physically as the instantaneous velocity at this point. maybe also a pict. The straight line  $T$  given parametrically as

$$T : c(t_0) + (t - t_0) \cdot c'(t_0)$$

is called the *tangent line to the curve  $c$*  at the point  $t_0$ . Unlike the tangent vector, the (unparametrized) tangent line  $T$  is independent of the parametrization of the curve  $c$ . The chain rule ensures that changing the parametrization leads to the same tangent vector, up to multiple.

**8.1.6. Partial derivatives.** If we look at the multivariate function  $f(x_1, \dots, x_n) : \mathbb{R}^n \rightarrow \mathbb{R}$  as at the function of one real variable  $x_i$  while the other variables are assumed constant, we can consider the derivative of this function. This is called the *partial derivative of the function  $f$*  with respect to  $x_i$ , and it is denoted as  $\frac{\partial f}{\partial x_i}$ ,  $i = 1, \dots, n$ , or (without referring to the particular function) as the operator  $\frac{\partial}{\partial x_i}$  on the functions.

More generally, for every function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and an arbitrary curve  $c : \mathbb{R} \rightarrow \mathbb{R}^n$ , their composition  $(f \circ c)(t) : \mathbb{R} \rightarrow \mathbb{R}$  can be considered. This composite function  $F = f \circ c$  expresses the behaviour of the function  $f$  along the curve





the derivatives at the point  $t = 1$  are 1, 2, 3. Therefore, the parametric equations of the tangent line are:

$$\begin{aligned} x &= c'_1(1)s + c_1(1) = t + 1, \\ y &= c'_2(1)s + c_2(1) = 2t + 1, \\ z &= c'_3(1)s + c_3(1) = 3t + 1. \end{aligned}$$

In order to get the implicit equations (which are not given canonically), we eliminate the parameter  $t$ , thereby obtaining:

$$\begin{aligned} 2x - y &= 1, \\ 3x - z &= 2. \end{aligned} \quad \square$$

**8.D.3.** Determine the tangent line  $p$  to the curve  $c(t) = (\ln t, \arctan t, e^{\sin(\pi t)})$  at the point  $t_0 = 1$ .



**8.D.4.** Find a point on the curve  $c(t) = (t^2 - 1, -2t^2 + 5t, t - 5)$  such that the tangent line passing through it is parallel to the plane  $\rho: 3x + y - z + 7 = 0$ .

**Solution.** The direction  $c'(t_0)$  of the curve  $c(t)$  in  $t_0$  has to be perpendicular to the normal of  $\rho$ , that is the scalar multiple of these two vectors is 0. The tangent vector in the point  $c(t)$  is  $(2t, -4t + 5, 1)$ , the normal vector of the plane  $\rho$  is  $(3, 1, -1)$  (just read off the coefficients by  $x, y$  and  $z$  in the equation of  $\rho$ . That is  $3 \cdot 2t + 1 \cdot (-4t + 5) + 1 \cdot 1 = 0$ , which gives  $[3, -18, -7]$ .  $\square$

**8.D.5.** Find the parametric equation of the tangent line of the curve given as the intersection of surfaces  $x^2 + y^2 + z^2 = 4$  and  $x^2 + y^2 - 2x = 0$  in the point  $[1, 1, \sqrt{2}]$ .

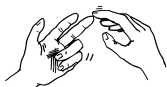
**Solution.**  $p = \{[1 - \sqrt{2}s, 1, \sqrt{2} + s]; s \in \mathbb{R}\}$ .  $\square$

**8.D.6. The set of differentiable functions.** We can notice that multivariate polynomials are differentiable on the whole of their domain. Similarly, the composition of a differentiable univariate function and a differentiable multivariate function leads to a differentiable multivariate function. For instance, the function  $\sin(x + y)$  is differentiable on the whole  $\mathbb{R}^2$ ;  $\ln(x + y)$  is a differentiable function on the set of points with  $x > y$  (an open half-plane, i. e., without the boundary line). The proofs of these propositions are left as an exercise on limit compositions.



**Remark.** Notation of partial derivatives. The partial derivative of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  in variables  $x_1, \dots, x_n$  with respect to the variable  $x_1$  will be denoted by both  $\frac{\partial f}{\partial x_1}$  and the shorter expression  $f_{x_1}$ . In the exercise part of the book, we will rather keep to the latter notation. On the other hand, the notation  $\frac{\partial f}{\partial x_1}$  better catches the fact that this is a derivative of  $f$  in the direction of the vector field  $\frac{\partial}{\partial x_1}$  (you will learn what a vector field is in paragraph 9.1.1).

**8.D.7.** Determine the domain of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2 \sqrt{y}$ . Calculate the partial derivatives where they are defined on this domain.



*c.* The simplest case is using parametrized straight lines  $c$  and choosing the lines  $c_i(t) = (x_1, \dots, x_i + t, \dots, x_n)$ , the derivative of  $f \circ c_i$  yields just the partial derivatives  $\frac{\partial f}{\partial x_i}$ . More generally, derivatives can be defined in any direction:

DIRECTIONAL AND PARTIAL DERIVATIVES

**Definition.**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  has derivative in the direction of a vector  $v \in \mathbb{R}^n$  at a point  $x \in E_n$  if and only if the derivative  $d_v f(x)$  of the composite mapping

$$t \mapsto f(x + tv)$$

exists at the point  $t = 0$ , i.e.

$$d_v f(x) = \lim_{t \rightarrow 0} \frac{1}{t} (f(x + tv) - f(x)).$$

The partial derivatives are the values  $\frac{\partial f}{\partial x_i} = d_{e_i} f$  where  $e_i$  are the elements of the standard basis of  $\mathbb{R}^n$ .

In other words, the directional derivative expresses the infinitesimal increment of the function  $f$  in the direction  $v$ .

For functions in the plane,

$$\frac{\partial}{\partial x} f(x, y) = \lim_{t \rightarrow 0} \frac{1}{t} (f(x + t, y) - f(x, y))$$

$$\frac{\partial}{\partial y} f(x, y) = \lim_{t \rightarrow 0} \frac{1}{t} (f(x, y + t) - f(x, y)).$$

Especially, the partial differentiation with respect to a given variable is just the casual one-variable differentiation while considering the other variables to be constants.

**8.1.7. The differential of a function.** Partial or directional derivatives are not always good enough to obtain a fair approximation of the behaviour of a function by linear expressions. There are three concerns for a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  there. First, the directional derivatives at a point  $x \in \mathbb{R}^n$  may not exist in all directions, although the partial derivatives are well defined. Second, the dependence of the directional derivatives  $d_v f(x)$  on the direction  $v$  need not be linear. Third, even if  $d_v f(x)$  is a linear mapping in the argument  $v$ , the function still may be not 'well behaved' around the point  $x$ .

As an example, consider the functions in the plane with coordinates  $(x, y)$  given by the formulae

$$g(x, y) = \begin{cases} 1 & \text{if } xy = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h(x, y) = \begin{cases} x & \text{if } y = 0 \\ y & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$k(x, y) = \begin{cases} x & \text{if } y = x^2 \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Both partial derivatives of  $g$  at  $(0, 0)$  exist and no other directional derivatives do, and  $g$  is even not continuous at the origin. The functions  $h$  and  $k$  are continuous at  $(0, 0)$  and  $h$  has all its directional derivatives at the origin equal zero,

**Solution.** The domain of the function in question in  $\mathbb{R}^2$  is the half-plane  $\{(x, y), y \geq 0\}$ . In order to determine the partial derivative with respect to a given variable, we consider the other variables to be constants in the formula that defines the function. Then, we simply differentiate the expression as a univariate function. We thus get:

$$f_x = 2xy \text{ a } f_y = \frac{1}{2} \frac{x^2}{\sqrt{y}}.$$

The partial derivatives exist at all points of the domain except for the boundary line  $y = 0$ .  $\square$

**8.D.8.** Determine the derivative of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^2yz$  at the point  $[1, -1, 2]$  in the direction  $v = (3, 2, -1)$ .

**Solution.** The directional derivative can be calculated in two ways. The first one is to derive it directly from the definition (see paragraph 8.1.6). The second one is to use the differential of the function; see 8.1.7 and theorem 8.1.8. Since the given function is a polynomial, it is differentiable on the whole  $\mathbb{R}^3$ .

Let us follow the definition:

$$\begin{aligned} f_v(x, y, z) &= \lim_{t \rightarrow 0} \frac{1}{t} [f(x + 3t, y + 2t, z - t) - f(x, y, z)] \\ &= \lim_{t \rightarrow 0} \frac{1}{t} [(x + 3t)^2(y + 2t)(z - t) - x^2yz] \\ &= \lim_{t \rightarrow 0} \frac{1}{t} [t(6xyz + 2x^2z - x^2y) + t^2(\dots)] \\ &= 6xyz + 2x^2z - x^2y. \end{aligned}$$

We have thus derived the derivative in the direction of the vector  $(3, 2, -1)$  as a function of three real variables which determine the point at which we are interested in the value of the derivative. Evaluating this for the desired point thus leads to  $f_v(1, -1, 2) = -7$ .

In order to compute the directional derivative from the differential of the function, we first have to determine the partial derivatives of the function:

$$f_x = 2xyz, \quad f_y = x^2z, \quad f_z = x^2y.$$

It follows from the note beyond theorem 8.1.8 that we can express

$$\begin{aligned} f_v(1, -1, 2) &= 3f_x(1, -1, 2) + 2f_y(1, -1, 2) + \\ &\quad + (-1)f_z(1, -1, 2) = \\ &= 3 \cdot (-4) + 2 \cdot 2 + (-1) \cdot (-1) = -7. \quad \square \end{aligned}$$

**8.D.9.** Determine the derivative of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = \frac{\cos(x^2y)}{z}$  at the point  $[0, 0, 2]$  in the direction of the vector  $(1, 2, 3)$ .

**Solution.** The domain of this function is  $\mathbb{R}^3$  except for the plane  $z = 0$ . The following calculations will be considered only on this domain. The function in question is differentiable at the point  $[0, 0, 2]$  (this follows from the note 8.D.6). We can determine the value of the examined directional derivative by 8.1.7, using partial derivatives.

First, we determine the partial derivatives of the given function (as we have already mentioned in exercise 8.D.7, in

except for the partial derivatives, which are equal to 1. In particular,  $d_v h(0, 0)$  is not a linear mapping in the argument  $v$ . More generally, consider a function  $f$  which, along the lines  $(r \cos \theta, r \sin \theta)$  with a fixed angle  $\theta$ , takes the values  $\alpha(\theta)r$ , where  $\alpha(\theta)$  is a periodic odd function of the angle  $\theta$ , with period  $2\pi$ . All of its directional derivatives  $d_v f$  at  $(0, 0)$  exist, yet these are not linear expressions depending on the directions  $v$  for general functions  $\alpha(\theta)$ . The graph of  $f$  can be visualized as a “deformed cone” and we can hardly hope for a good linear approximation at its vertex.

Finally,  $k$  has all directional derivatives zero, i.e.  $d_v h(0) = 0$  for all directions  $v$ , which is a linear dependence on  $v \in \mathbb{R}^2$ . But still, the zero mapping is a very bad approximation of  $k$  along the parabola  $y = x^2$ . Check all these claims in detail yourselves!

Therefore, we imitate the case of univariate functions as thoroughly as possible, and avoid such a pathological behaviour of functions directly by defining and using the concept of differential:

DIFFERENTIAL OF A FUNCTION

**Definition.** A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  has got the *differential* at a point  $x$  if and only if all of the following three conditions hold:

- (1) the directional derivatives  $d_v f(x)$  at the point  $x$  exist for all vectors  $v \in \mathbb{R}^n$ ,
- (2)  $d_v f(x)$  is linearly dependent on the argument  $v$ ,
- (3)  $\lim_{v \rightarrow 0} \frac{1}{\|v\|} (f(x + v) - f(x) - d_v f(x)) = 0$ .

The linear expression  $d_v f$  (in a vector variable  $v$ ) is then called the *differential  $df$  of the function  $f$*  evaluated at the increase  $v$ .

In words, it is required that the behaviour of the function  $f$  at the point  $x$  is well approximated by linear functions of increments of the variable quantities.

It follows directly from the definition of directional derivatives that the differential can be defined solely by the property (3). If there is a linear form  $df(x)$  such that the increments  $v$  at the point  $x$  satisfy the property (3) with  $d_v f(x) = df(x)(v)$ , then  $df(x)(v)$  is apparently just the directional derivative of the function  $f$  at the point  $x$ , so the properties (1) and (2) are automatically satisfied.

Notice, that in dimension one, the only linear functions are multiplications by constant numbers, and if such a number satisfying (3) exists, we call it the derivative  $f'(x)$  in the point  $x$ . Then the first two properties automatically hold true and thus we did not have to distinguish the derivative and the differential there.

**8.1.8.** Examine what can be said about the differential of a function  $f(x, y)$  in the plane, supposing both partial derivatives  $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$  exist and are continuous in a neighbourhood of a point  $(x_0, y_0)$ . To this purpose, consider any smooth curve  $t \mapsto (x(t), y(t))$  with  $x_0 = x(0), y_0 = y(0)$ .



order to determine the partial derivative with respect to  $x$ , we differentiate it as a univariate function (in  $x$ ) and use the chain rule; similarly for other partial derivatives):

$$f_x = -\frac{2xy \sin(x^2y)}{z}, \quad f_y = -\frac{x^2 \sin(x^2y)}{z},$$

$$f_z = -\frac{\cos(x^2y)}{z^2}.$$

Evaluating the expression gives

$$f_x(0, 0, 2) + 2 \cdot f_y(0, 0, 2) + 3 \cdot f_z(0, 0, 2)$$

$$= 1 \cdot 0 + 2 \cdot 0 + 3 \cdot \left(-\frac{1}{4}\right) = -\frac{3}{4}. \quad \square$$

**8.D.10.** Having a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with differential  $df(x)$  and a point  $x \in \mathbb{R}^n$ , determine a unit direction  $v \in \mathbb{R}^n$  in which the directional derivative  $d_v(x)$  is maximal.

**Solution.** According to the note beyond theorem 8.1.5, we are maximizing the function  $f_v(x) = v_1 f_{x_1}(x) + v_2 f_{x_2}(x) + \dots + v_n f_{x_n}(x)$  in dependence on the variables  $v_1, \dots, v_n$  which are bound by the condition  $v_1^2 + \dots + v_n^2 = 1$ . We have already solved this type of problem in chapter 3, when we talked about linear optimization (viz 3.A.5). The value  $f_v(x)$  can be interpreted as the scalar product of the vectors  $(f_{x_1}, \dots, f_{x_n})$  and  $(v_1, \dots, v_n)$ . And this product is maximal if the vectors are of the same direction. The vector  $v$  can thus be obtained by normalizing the vector  $(f_{x_1}, \dots, f_{x_n})$ . In general, we say the the function grows maximally in the direction  $(f_{x_1}, \dots, f_{x_n})$ . Then, this vector is called the gradient of the function  $f$ . In paragraph 8.1.26, we will recall this idea and go into further details.  $\square$

Counting the differential of a function is technicly very easy, just plug into the definition.

**8.D.11.** Find the differential of a function  $f$  in a point  $P$ :

- i)  $f(x, y) = \arctan \frac{x+y}{1-xy}, P = [\sqrt{3}, 1]$
- ii)  $f(x, y) = \arcsin \frac{x}{\sqrt{x^2+y^2}}, P = [1, \sqrt{3}]$ .
- iii)  $f(x, y) = xy + \frac{x}{y}, P = [1, 1]$ .

**Solution.**

- i)  $df(\sqrt{3}, 1) = \frac{1}{4}dx + \frac{1}{2}dy,$
- ii)  $df(1, \sqrt{3}) = \frac{\sqrt{3}}{4}dx - \frac{1}{4}dy,$
- iii)  $df(1, 1) = 2dx.$

$\square$

Let us realize, that differential of a function is a linear map:

**8.D.12.** Count the differential of the function  $f(x, y, z) = 2^x \sin y \arctan z$  in the point  $[-4, \frac{\pi}{2}, 0]$  evaluated on  $dx = 0.05, dy = 0.06,$  and  $dz = 0.08.$

**Solution.**  $df(-4, \frac{\pi}{2}, 0) = 0dx + 0dy + \frac{1}{16}dz = 0.005.$   $\square$

The differential thus can be used to approximate the values of a function.

The idea is to use the mean value theorem for univariate functions for differences of function values, where only one of the variables changes:  $f(x, y) - f(x_0, y) = \frac{\partial f}{\partial x}(x_1, y)(x - x_0)$  for suitable  $x_1$  between  $x_0$  and  $x$ .

Apply this in both summands of the following expression separately, to obtain

$$\frac{1}{t}(f(x(t), y(t)) - f(x_0, y_0)) =$$

$$\frac{1}{t}(f(x(t), y(t)) - f(x_0, y(t))) + \frac{1}{t}(f(x_0, y(t)) - f(x_0, y_0))$$

$$= \frac{1}{t}(x(t) - x_0) \frac{\partial f}{\partial x}(x(\xi), y(t)) + \frac{1}{t}(y(t) - y_0) \frac{\partial f}{\partial y}(x_0, y(\eta))$$

for suitable numbers  $\xi$  and  $\eta$  between 0 and  $t$ . Indeed, by exploiting that the curve  $(x(t), y(t))$  is differentiable, there must be such values  $\xi$  and  $\eta$ .

Especially, for every sequence of numbers  $t_n$  converging to zero, the corresponding sequences of numbers  $\xi_n$  and  $\eta_n$  also converge to zero (by the squeeze theorem for three limits) and they all satisfy the above equality.

If  $t$  converges to 0, the continuity of the partial derivatives, together with the test for convergence of functions using subsequences of the input values (cf. 5.2.15), as well as the properties of the limits of sums and products of functions (cf. Theorem 5.2.13) imply

$$\frac{d}{dt} f(x(t), y(t))|_{t=0} = x'(0) \frac{\partial f}{\partial x}(x_0, y_0) + y'(0) \frac{\partial f}{\partial y}(x_0, y_0),$$

which is a pleasant extension of the theorem on differentiation of composite functions of one variable for the case  $f \circ c$ .

Of course, with the special choice of parametrized straight lines with direction vector  $v = (\xi, \eta)$ ,

$$(x(t), y(t)) = (x_0 + t\xi, y_0 + t\eta),$$

the calculation leads to the derivative in the direction  $v = (\xi, \eta)$  and the equality

$$d_v f(x_0, y_0) = \frac{\partial f}{\partial x}(x_0, y_0)\xi + \frac{\partial f}{\partial y}(x_0, y_0)\eta.$$

This formula can be expressed in a neat way to describe coordinate expressions of linear functions on vector spaces:

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy,$$

where  $dx$  stands for the differential of the function  $(x, y) \mapsto x$ , i.e.  $dx(v) = \xi$ , and similarly for  $dy$ . In other words, the directional derivative  $d_v f$  is a linear function  $\mathbb{R}^n \rightarrow \mathbb{R}$  on the increments, with coordinates given by the partial derivatives.

Now we could similarly prove that the assumption of continuous partial derivatives at a given point guarantees the approximation property of the differential as well. In particular, note that the computation for  $f \circ c$  above excluded phenomena like the function  $k(x, y)$  above (there  $d_v k(0, 0) = 0$ , but the derivative along the curve  $(t, t^2)$  was one). We shall better do this for the general multivariate functions straightaway.

**8.D.13.** Approximate  $\sqrt{2.98^2 + 4.05^2}$  with the use of differential (and not with a calculator).

**Solution.** We use the differential of the function  $f(x, y) = \sqrt{x^2 + y^2}$  in  $[3, 4]$ . Then  $f'_x = \frac{x}{\sqrt{x^2 + y^2}}$ ,  $f'_y = \frac{y}{\sqrt{x^2 + y^2}}$ , thus

$$\begin{aligned} \sqrt{2.98^2 + 4.05^2} &\doteq f(3^2, 4^2) + df(2.98 - 3, 4.05 - 4) \\ &= \sqrt{3^2 + 4^2} + \frac{3}{\sqrt{3^2 + 4^2}}(-0.02) + \frac{4}{\sqrt{3^2 + 4^2}}(0.05) \\ &= 5 - \frac{0,06}{5} + \frac{0,2}{5} = 5,028. \quad \square \end{aligned}$$

**8.D.14.** With the help of a differential calculate

- i)  $\arctan \frac{1,02}{0,95}$ ,
- ii)  $\ln(0,97^2 + 0,05^2)$ ,
- iii)  $\arcsin \frac{0,48}{1,05}$
- iv)  $1,04^{2,02}$ .

**8.D.15.** What is approximately the change (in  $cm^3$ ) in the volume of the cone with a base radius  $r = 10$  cm and height  $h = 10$  cm, if we increase the radius by 5 mm and we decrease the height by 5 mm?

**Solution.** The volume is (as a function of the radius  $r$  and a height  $h$ )  $V(r, h) = \frac{1}{3}\pi r^2 h$ . The change is approximately given by the differential of  $V$  in  $[10, 10]$  evaluated on  $dr = 10.5 - 10 = 0.5$  and  $dh = 9.5 - 10 = -0.5$ . We get  $\frac{50}{3}\pi cm^3$ .  $\square$

**8.D.16.** Find the tangent plane to the graph of a function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  in a point  $P = [x_0, y_0, f(x_0, y_0)]$ :

- i)  $f(x, y) = \sqrt{1 - x^2 - y^2}$ ,  $P = [x_0, y_0, z_0] = [\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, ?]$ .
- ii)  $f(x, y) = e^{x^2 + y^2}$ ,  $P = [x_0, y_0, z_0] = [0, 0, ?]$ ,
- iii)  $f(x, y) = x^2 + xy + 2y^2$ ,  $P = [x_0, y_0, z_0] = [1, 1, ?]$ ,
- iv)  $f(x, y) = \arctan \frac{y}{x}$ ,  $P = [x_0, y_0, z_0] = [1, -1, ?]$ .

**Solution.**

i)  $f(x_0, y_0) = \sqrt{1 - \frac{1}{3} - \frac{1}{3}} = \frac{1}{\sqrt{3}}$ , thus  $z_0 = \frac{1}{\sqrt{3}}$ . Further  $f'_x = -\frac{x}{\sqrt{1 - x^2 - y^2}}$ ,  $f'_y = -\frac{y}{\sqrt{1 - x^2 - y^2}}$ , thus  $f'_x(x_0, y_0) = -\frac{1/\sqrt{3}}{1/\sqrt{3}} = -1$ ,  $f'_y(x_0, y_0) = -\frac{1/\sqrt{3}}{1/\sqrt{3}} = -1$ . The equation of a tangent plane in  $[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]$  is

$$z = \frac{1}{\sqrt{3}} - (x - \frac{1}{\sqrt{3}}) - (y - \frac{1}{\sqrt{3}}), \text{ or } x + y + z = \sqrt{3},$$

- ii)  $z_0 = 1, z = 1$ ,
- iii)  $z_0 = 4, 3x + 5y - z = 4$ ,
- iv)  $z_0 = -\frac{\pi}{4}, x + y - 2z = \frac{\pi}{2}$ .

$\square$

**8.1.9.** The following theorem provides a crucial and very useful observation.

CONTINUITY OF PARTIAL DERIVATIVES

**Theorem.** Let  $f : E_n \rightarrow \mathbb{R}$  be a function of  $n$  variables with continuous partial derivatives in a neighbourhood of the point  $x \in E_n$ . Then its differential  $df$  at the point  $x$  exists and its coordinate expression is given by the formula

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \dots + \frac{\partial f}{\partial x_n} dx_n.$$

**PROOF.** This theorem can be derived analogously to the procedure described above, for the case  $n = 2$ . Care is needed in details to finish the reasoning about the approximation property. As above, consider a curve

$c(t) = (c_1(t), \dots, c_n(t))$ ,  $c(0) = (0, \dots, 0)$  and a point  $x \in \mathbb{R}^n$ , and express the difference  $f(x + c(t)) - f(x)$  for the composite function  $f(c(t))$  as follows:

$$\begin{aligned} &f(x_1 + c_1(t), \dots, x_n + c_n(t)) - f(x_1, x_2 + c_2(t), \dots) \\ &+ f(x_1, x_2 + c_2(t), \dots) - f(x_1, x_2, \dots, x_n + c_n(t)) \\ &\vdots \\ &+ f(x_1, x_2, \dots, x_n + c_n(t)) - f(x_1, x_2, \dots, x_n). \end{aligned}$$

Now, apply the mean value theorem to all of the  $n$  summands, obtaining (similarly to the case of two variables)

$$\begin{aligned} &(c_1(t) - c_1(0)) \frac{\partial f}{\partial x_1}(x_1 + c_1(\theta_1), \dots, x_n + c_n(t)) \\ &+ (c_2(t) - c_2(0)) \frac{\partial f}{\partial x_2}(x_1, x_2 + c_2(\theta_2), \dots, x_n + c_n(t)) \\ &\vdots \\ &+ (c_n(t) - c_n(0)) \frac{\partial f}{\partial x_n}(x_1, x_2, \dots, x_n + c_n(\theta_n)), \end{aligned}$$

for appropriate values  $\theta_i$ ,  $0 \leq \theta_i \leq t$ . This is a finite sum, so the same reasoning as in the case of two variables verifies that

$$\frac{d}{dt} f(x + c(t))|_{t=0} = c'_1(0) \frac{\partial f}{\partial x_1}(x) + \dots + c'_n(0) \frac{\partial f}{\partial x_n}(x).$$

The special choice of the curves  $c(t) = x + tv$  for a directional vector  $v$  verifies the statement about existence and linearity of the directional derivatives at  $x$ .

Finally, apply the mean value theorem in the same way to the difference

$$\begin{aligned} f(x + v) - f(x) &= d_v f(x + \theta v) \\ &= v_1 \frac{\partial f}{\partial x_1}(x + \theta v) + \dots + v_n \frac{\partial f}{\partial x_n}(x + \theta v) \end{aligned}$$

**8.D.17.** Find all points on the conic  $k : x^2 + 3y^2 - 2x + 6y - 8 = 0$  such that the normal of the conic is parallel to the  $y$  axis. For each point write the equation of the tangent in the point.

**Solution.** The normal to  $k$  in a point is parallel to the  $y$  axis iff the tangent line in the point is parallel to the  $x$  axis. The normal to  $k$  in  $[x_0, y_0] \in k$  is parallel to  $y$  axis iff one of the tangents to  $k$  in  $[x_0, y_0]$  is parallel to  $x$  axis, and this happens iff  $y'(x_0) = 0$ , where  $y$  is a function given implicitly by  $k$  in a neighborhood of  $[x_0, y_0]$ . Derivation of the equation of  $k$  gives  $2x + 6yy' - 2 + 6y' = 0$ , that is  $y' = \frac{1-x}{3(1+y)}$ . Thus  $y(x_0)' = 0$ , iff  $x_0 = 1$ . Substituting to the equation of  $k$  we get  $1 + 3y_0^2 - 2 + 6y_0 - 8 = 0$ , thus  $y_0 = 1$  or  $y_0 = -3$ . The sought points are  $[1, 1]$ , resp.  $[1, -3]$ , the equations of tangents in the points are  $y = 1$ , resp.  $y = -3$ .  $\square$

**8.D.18.** On the conic given by the equation  $3x^2 + 6y^2 - 3x + 3y - 2 = 0$  find all points where the normal to the conic is parallel with the line  $y = x$ . For each point give the equation of the tangent in the point.  $\circ$

**8.D.19.** On the conic given by the equation  $x^2 + xy + 2y^2 - x + 3y - 54 = 0$  find all points where the normal to the conic is parallel to the  $x$  axis. For each point give the equation of the tangent in the point.  $\circ$

**8.D.20.** On the graph of the function  $u(x, y, z) = x\sqrt{y^2 + z^2}$  find all points where the tangent plane is parallel to the plane  $x + y - z - u = 0$ .  $\circ$

**8.D.21.** Find the points on to the ellipsoid  $x^2 + 2y^2 + z^2 = 1$ , where the tangent planes are parallel to  $x - y + 2z = 0$ .

**Solution.** The equation of the tangent plane is determined by the partial derivatives of  $z = z(x, y)$  given implicitly by the equation  $x^2 + 2y^2 + z^2 = 1$  of the ellipsoid. The normal vector in  $[x_0, y_0, z_0]$  is  $(z'_x(x_0, y_0), z'_y(x_0, y_0), -1)$ . This vector has to be parallel to the normal  $(1, -1, 2)$  of the plane, thus  $(-2z'_x(x_0, y_0), -2z'_y(x_0, y_0), 2) = (1, -1, 2)$ . which yields  $2x_0 = z_0, 4y_0 = -z_0$  and after substituting to the ellipsoid's equation we get the sought points:  $[\frac{2}{\sqrt{22}}, -\frac{1}{\sqrt{22}}, \frac{4}{\sqrt{22}}]$  and  $[-\frac{2}{\sqrt{22}}, +\frac{1}{\sqrt{22}}, -\frac{4}{\sqrt{22}}]$ .

**Another solution.** It is useful to realize, that the normal vector in  $[x_0, y_0, z_0]$  of the surface given implicitly by  $F(x, y, z) = 0$  is the vector  $(F'_x(x_0, y_0, z_0), F'_y(x_0, y_0, z_0), F'_z(x_0, y_0, z_0))$ .  $\square$

**8.D.22.** Determine whether the tangent plane to the graph of the function  $f : \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R}, f(x, y) = x \cdot \ln(y)$  at the point  $[1, \frac{1}{e}]$  goes through the point  $[1, 2, 3] \in \mathbb{R}^3$ .

**Solution.** First of all, we calculate the partial derivatives:  $f_x(x, y) = \ln(y), f_y(x, y) = \frac{x}{y}$ ; their values at the point  $[1, \frac{1}{e}]$  are  $-1, e$ ; further  $f(1, \frac{1}{e}) = -1$ . Therefore, the equation of the tangent plane is

$$z = f\left(1, \frac{1}{e}\right) + f_x\left(1, \frac{1}{e}\right)(x - 1) + f_y\left(1, \frac{1}{e}\right)\left(y - \frac{1}{e}\right) = -1 - x + ey.$$

with an appropriate  $\theta, 0 \leq \theta \leq 1$ , where the latter equality holds according to the formula for directional derivatives derived above, for sufficiently small  $v$ 's.

Since all the partial derivatives are continuous at the point  $x$ , for an arbitrarily small  $\varepsilon > 0$ , there is a  $\delta$ -neighbourhood  $U$  of the origin in  $\mathbb{R}^n$  such that for  $w \in U$ , all partial derivatives  $\frac{\partial f}{\partial x_i}(x + w)$  differ from  $\frac{\partial f}{\partial x_i}(x)$  by less than  $\varepsilon$ . Hence the estimate

$$\begin{aligned} \frac{1}{\|w\|} \|f(x + w) - f(x) - d_w f(x)\| &\leq \\ &\leq \frac{1}{\|w\|} \|f(x + w) - f(x) - d_w f(x + \theta w)\| + \\ &\quad \frac{1}{\|w\|} \|d_w f(x + \theta w) - d_w f(x)\| \\ &= \frac{1}{\|w\|} \|w_1 \left(\frac{\partial f}{\partial x_1}(x + \theta w) - \frac{\partial f}{\partial x_1}(x)\right) + \dots \\ &\quad + w_n \left(\frac{\partial f}{\partial x_n}(x + \theta w) - \frac{\partial f}{\partial x_n}(x)\right)\| \\ &\leq \frac{n}{\|w\|} \|w\| \varepsilon, \end{aligned}$$

where  $\theta$  is the parameter for which the expression on the second line vanishes. Thus, the approximation property of the differential is satisfied as well.  $\square$

The approximation property of the differential can be written as

$$f(x + v) = f(x) + df(x)(v) + \alpha(v),$$

where the function  $\alpha(v)$  satisfies  $\lim_{v \rightarrow 0} \frac{\alpha(v)}{\|v\|} = 0$ , i.e.  $\alpha(v) = o(\|v\|)$  in the asymptotic terminology introduced in 6.1.12 on the page 520.

**8.1.10. A plane tangent to the graph of a function.** The linear approximation of the function behaviour by its differential can be expressed in terms of its graph, similarly to the case of univariate functions. We work with hyperplanes instead of tangent lines.

In the case of a function on  $E_2$  and a fixed point  $(x_0, y_0) \in E_2$ , consider the plane in  $E_3$  given by the equation on the three coordinates  $(x, y, z)$ :

$$\begin{aligned} z &= f(x_0, y_0) + df(x_0, y_0)(x - x_0, y - y_0) \\ &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0). \end{aligned}$$

It is already seen that the increase of the function values of a differentiable function  $f : E_n \rightarrow \mathbb{R}$  at points  $x + tv$  and  $x$  is always expressed in terms of the directional derivative  $d_v f$  at a suitable point between them. Therefore, this is the only plane containing the point  $(x_0, y_0)$  with the property that all derivatives, and so the tangent lines of all curves

$$c(t) = (x(t), y(t), f(x(t), y(t)))$$

lie in this plane, too. It is called the *tangent plane* to the graph of the function  $f$ .



The given point does not satisfy this equation, so it does not lie in the tangent plane.  $\square$

**8.D.23.** Determine the parametric equation of the tangent line to the intersection of the graphs of the functions  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2 + xy - 6$ ;  $g : \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R}$ ,  $g(x, y) = x \cdot \ln(y)$  at the point  $[2, 1]$ .

**Solution.** The tangent line to the intersection is the intersection of the tangent planes at the given point. The plane that is tangent to the graph of  $f$  and goes through the point  $[2, 1]$  is

$$z = f(2, 1) + f_x(2, 1)(x - x_0) + f_y(2, 1)(y - y_0) = 5x + 2y - 12.$$

The tangent plane to the graph of  $g$  is then

$$z = f(2, 1) + g_x(x, y)(2, 1)(x - x_0) + g(x, y)_y(2, 1)(y - y_0) = 2y - 2.$$

The intersection line of these two planes is given parametrically as  $[2, t, 2t - 2]$ ,  $t \in \mathbb{R}$ .

**Another solution.** The normal to the surface given by the equation  $f(x, y, z) = 0$  at the point  $b = [2, 1, 0]$  is  $(f_x(b), f_y(b), f_z(b)) = (5, 2, -1)$ ; the normal to the surface given by  $g(x, y, z) = 0$  at the same point is  $(0, 2, -1)$ . The tangent line is perpendicular to both normals; we can thus obtain a vector it is parallel to as the vector product of the normals, which is  $(0, 5, 10)$ . Since the tangent line goes through the point  $[2, 1, 0]$ , its parametric equation is  $[2, 1 + t, 2t]$ ,  $t \in \mathbb{R}$ .  $\square$

**8.D.24. V.** ypočtete všechny parciální derivace prvního a druhého řádu funkce  $f(x, y, z) = x^{\frac{y}{z}}$ .

**Solution.**  $f'_x = \frac{y}{z} x^{\frac{y}{z}-1}$ ,  $f'_y = x^{\frac{y}{z}} \ln x \cdot \frac{1}{z}$ ,  $f'_z = x^{\frac{y}{z}} \ln x \cdot \frac{-y}{z^2}$ ,  
 $f''_{xx} = \frac{y}{z} (\frac{y}{z} - 1) x^{\frac{y}{z}-2}$ ,  $f''_{yy} = x^{\frac{y}{z}} \ln^2 x \cdot \frac{1}{z^2}$ ,  
 $f''_{zz} = x^{\frac{y}{z}} \ln^2 x \cdot \frac{y^2}{z^4} + x^{\frac{y}{z}} \ln x \cdot \frac{2y}{z^3}$ ,  
 $f''_{xy} = \frac{1}{z} x^{\frac{y}{z}-1} + \frac{y}{z} x^{\frac{y}{z}-1} \ln x \cdot \frac{1}{z}$ ,  $f''_{xz} = \frac{-y}{z^2} x^{\frac{y}{z}-1} + \frac{y}{z} x^{\frac{y}{z}-1} \ln x \cdot \frac{-y}{z^2}$ ,  $f''_{yz} = x^{\frac{y}{z}} \ln^2 x \cdot \frac{-y}{z^3} + x^{\frac{y}{z}} \ln x \cdot \frac{-1}{z^2}$ .

**8.D.25.** Find all first and second order partial derivatives of  $z = f(x, y)$  in  $[1, \sqrt{2}, 2]$  defined in a neighborhood of the point by  $x^2 + y^2 + z^2 - xz - \sqrt{2}yz = 1$ .  $\circ$

**8.D.26.** Find all first and second order partial derivatives of  $z = f(x, y)$  in  $[-2, 0, 1]$  defined in a neighborhood of the point by  $2x^2 + 2y^2 + z^2 + 8xz - z + 8 = 0$ .  $\circ$

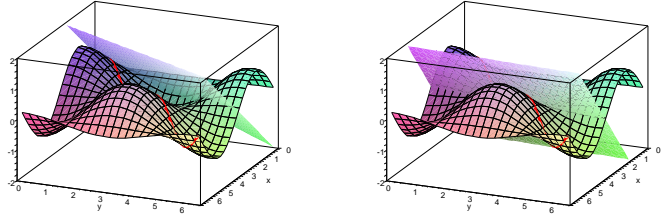
**8.D.27.** Determine all second partial derivatives of the function  $f$  given by  $f(x, y, z) = \sqrt{xy} \ln z$ .

**Solution.** First, we determine the domain of the given function: the argument of the square root must be non-negative, and the argument of the natural logarithm must be positive. Therefore,  $Df = \{(x, y, z) \in \mathbb{R}^3, (z \geq 1 \& (xy > 0)) \vee (0 < z < 1) \& (xy < 0)\}$ .

Two tangent planes to the graph of the function

$$f(x, y) = \sin(x) \cos(y)$$

are shown in the illustration. The diagonal line is the image of the curve  $c(t) = (t, t, f(t, t))$ .



For the case of functions of  $n$  variables, the tangent plane is defined as an analogy to the tangent plane to a surface in the three-dimensional space. Instead of being overwhelmed by many indices, it is useful to recall affine geometry, where hyperplanes can be used, see paragraph 4.1.3.



#### TANGENT (HYPER)PLANES

**Definition.** A *tangent hyperplane* to the graph of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  at a point  $x \in \mathbb{R}^n$  is the hyperplane containing the point  $(x, f(x))$  with the modelling vector space which is the graph of the linear mapping  $df(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ , i.e. the differential at the point  $x \in E_n$ .

The definition takes advantage of the fact that the directional derivative  $d_v f$  is given by the increment in the tangent (hyper)plane corresponding to the increment  $v$ .

Many analogies with the univariate functions follow from the latter fact. In particular, a differentiable function  $f$  on  $E_n$  has zero differential at a point  $x \in E_n$  if and only if its composition with any curve going through this point has a stationary point there, i.e., is neither increasing, nor decreasing in the linear approximation.

In other words, the tangent plane at such a point is parallel to the hyperplane of the variables (i.e., its modelling space is  $E_n \subset E_{n+1}$ , having added the last coordinate set to zero). Of course, this does not mean that  $f$  should have a local extremum at such a point. Just as in the case of univariate functions, this depends on the values of higher derivatives. But it is a necessary condition to the existence of extrema.

**8.1.11. Derivatives of higher orders.** The operation of differentiation can be iterated similarly to the case of univariate functions. This time, choose new directions for each iteration.



Fix an increment  $v \in \mathbb{R}^n$ . The enumeration of the differentials at this argument defines a (differential) operation on differentiable functions  $f : E_n \rightarrow \mathbb{R}$

$$f \mapsto d_v f = df(v),$$

and the result is again a function  $df(v) : E_n \rightarrow \mathbb{R}$ . If this function is differentiable as well, repeat this procedure with

Now, we calculate the first partial derivatives with respect to each of the three variables:

$$f_x = \frac{y \ln(z)}{2\sqrt{xy \ln(z)}}, f_y = \frac{x \ln(z)}{2\sqrt{xy \ln(z)}}, f_z = \frac{xy}{2z\sqrt{xy \ln(z)}}.$$

Each of these three partial derivatives is again a function of three variables, so we can consider (first) partial derivatives of these functions. Those are the second partial derivatives of the function  $f$ . We will write the variable with respect to which we differentiate as a subscript of the function  $f$ .

$$\begin{aligned} f_{xx} &= -\frac{y^2 \ln^2 z}{4(xy \ln z)^{\frac{3}{2}}}, \\ f_{xy} &= -\frac{xy \ln^2 z}{4(xy \ln z)^{\frac{3}{2}}} + \frac{\ln z}{2\sqrt{xy \ln z}}, \\ f_{xz} &= -\frac{xy^2 \ln z}{4z(xy \ln z)^{\frac{3}{2}}} + \frac{y}{2z\sqrt{xy \ln z}}, \\ f_{yy} &= -\frac{x^2 \ln^2 z}{4(xy \ln z)^{\frac{3}{2}}}, \\ f_{yz} &= -\frac{x^2 y \ln z}{4z(xy \ln z)^{\frac{3}{2}}} + \frac{x}{2z\sqrt{xy \ln z}}, \\ f_{zz} &= -\frac{x^2 y^2}{4z^2(xy \ln z)^{\frac{3}{2}}} - \frac{xy}{2z^2\sqrt{xy \ln z}}. \end{aligned}$$

By the theorem about interchangeability of partial derivatives (see 8.1.12), we know that  $f_{xy} = f_{yx}$ ,  $f_{xz} = f_{zx}$ ,  $f_{yz} = f_{zy}$ . Therefore, it suffices to compute the mixed partial derivatives (the word “mixed” means that we differentiate with respect to more than one variable) just for one order of differentiation.  $\square$

### E. Taylor polynomials

**8.E.1.** Write the second-order Taylor expansion of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = \ln(x^2 + y^2 + 1)$  at the point  $[1, 1]$ .

**Solution.** First, we compute the first partial derivatives:

$$f_x = \frac{2x}{x^2 + y^2 + 1}, f_y = \frac{2y}{x^2 + y^2 + 1},$$

then the Hessian:

$$Hf(x, y) = \begin{pmatrix} \frac{2y^2 - 2x^2 + 2}{(x^2 + y^2 + 1)^2} & -\frac{4xy}{(x^2 + y^2 + 1)^2} \\ -\frac{4xy}{(x^2 + y^2 + 1)^2} & \frac{2x^2 - 2y^2 + 2}{(x^2 + y^2 + 1)^2} \end{pmatrix}.$$

The value of the Hessian at the point  $[1, 1]$  is

$$\begin{pmatrix} \frac{2}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{2}{9} \end{pmatrix}.$$

another increment, and so on. In particular, work with iterations of partial derivatives. For *second-order partial derivatives*, write

$$\left( \frac{\partial}{\partial x_j} \circ \frac{\partial}{\partial x_i} \right) f = \frac{\partial^2}{\partial x_i \partial x_j} f = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

In the case of the repeated choice  $i = j$ , write

$$\left( \frac{\partial}{\partial x_i} \circ \frac{\partial}{\partial x_i} \right) f = \frac{\partial^2}{\partial x_i^2} f = \frac{\partial^2 f}{\partial x_i^2}.$$

Proceed in the same way with further iterations and talk about *partial derivatives of order  $k$*

$$\frac{\partial^k f}{\partial x_{i_1} \dots \partial x_{i_k}}.$$

More generally, one can iterate (assuming the function is sufficiently differentiable) any directional derivatives; for instance,  $d_v \circ d_w f$  for two fixed increments  $v, w \in \mathbb{R}^n$ .

#### $k$ -TIMES DIFFERENTIABLE FUNCTIONS

**Definition.** A function  $f : E_n \rightarrow \mathbb{R}$  is  *$k$ -times (continuously) differentiable* at a point  $x$  if and only if all its partial derivatives up to order  $k$  (inclusive) exist in a neighbourhood of the point  $x$  and are continuous at this point.

$f$  is  *$k$ -differentiable* if it is  $k$ -times (continuously) differentiable at all points of its domain.

From now on, the work is with continuously differentiable functions unless explicitly stated otherwise.

To show the basic features of the higher derivatives in the simplest form, work in the plane  $E_2$ . In the plane as well as in the space, iterated derivatives are often denoted by mere indices referring to the variable names, for example:

$$f_x = \frac{\partial f}{\partial x}, f_{xx} = \frac{\partial^2 f}{\partial x^2}, f_{xy} = \frac{\partial^2 f}{\partial x \partial y}, f_{yx} = \frac{\partial^2 f}{\partial y \partial x}.$$

Supposing the second-order partial derivatives are continuous at the point  $x$  (and exist in its neighborhood), we show that the partial derivatives commute. That is, the order in which differentiation is carried out does not matter.



By our assumption, the limits

$$\begin{aligned} f_{xy}(x, y) &= \lim_{t \rightarrow 0} \frac{1}{t} (f_x(x, y + t) - f_x(x, y)) \\ &= \lim_{t \rightarrow 0} \frac{1}{t} \left( \lim_{s \rightarrow 0} \frac{1}{s} (f(x + s, y + t) - f(x, y + t)) \right. \\ &\quad \left. - (f(x + s, y) - f(x, y)) \right) \end{aligned}$$

exist. However, since the limits can be expressed by any choice of values  $t_n \rightarrow 0$  and  $s_n \rightarrow 0$  and the limits of the corresponding sequences, the second derivative can be expressed

Altogether, we get that the second-order Taylor expansion at the point  $[1, 1]$  is

$$\begin{aligned} T_2(x, y) &= f(1, 1) + f_x(1, 1)(x - 1) + f_y(1, 1)(y - 1) \\ &\quad + \frac{1}{2}(x - 1, y - 1)Hf(1, 1)\begin{pmatrix} x - 1 \\ y - 1 \end{pmatrix} \\ &= \ln(3) + \frac{2}{3}(x - 1) + \frac{2}{3}(y - 1) + \frac{1}{9}(x - 1)^2 \\ &\quad - \frac{4}{9}(x - 1)(y - 1) + \frac{1}{9}(y - 1)^2 \\ &= \frac{1}{9}(x^2 + y^2 + 8x + 8y - 4xy - 14) + \ln(3). \end{aligned}$$

□

**Remark.** In particular, we can see that the second-order Taylor expansion of an arbitrary differentiable function at a given point is a second-order polynomial.

**8.E.2.** Determine the second-order Taylor polynomial of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,  $f(x, y) = xy \cos y$  at the point  $[\pi, \pi]$ . Decide whether the tangent plane to the graph of this function at the point  $[\pi, \pi, f(\pi, \pi)]$  goes through the point  $[0, \pi, 0]$ .

**Solution.** As in the above exercises, we find out that

$$T(x, y) = \frac{1}{2}\pi^2 y^2 - xy - \pi^3 y + \frac{1}{2}\pi^4.$$

The tangent plane to the graph of the given function at the point  $[\pi, \pi]$  is given by the first-order Taylor polynomial at the point  $[\pi, \pi]$ ; its general equation is thus

$$z = -\pi y - \pi x + \pi^2,$$

and this equation is satisfied by the given point  $[0, \pi, 0]$ . □

**8.E.3.** Determine the third-order Taylor polynomial of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^3 y + xz^2 + xy + 1$  at the point  $[0, 0, 0]$ . ○

**8.E.4.** Determine the second-order Taylor polynomial of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2 \sin y + y^2 \cos x$  at the point  $[0, 0]$ . Decide whether the tangent plane to the graph of this function at the point  $[0, 0, 0]$  goes through the point  $[\pi, \pi, \pi]$ . ○

**8.E.5.** Determine the second-order Taylor polynomial of the function  $\ln(x^2 y)$  at the point  $[1, 1]$ . ○

**8.E.6.** Determine the second-order Taylor polynomial of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,

$$f(x, y) = \tan(xy + y)$$

at the point  $[0, 0]$ . ○

### F. Extrema of multivariate functions

**8.F.1.** Determine the stationary points of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2 y + y^2 x - xy$  and decide which of these points are local extrema and of which type.

**Solution.** The first derivatives are  $f_x = 2xy + y^2 - y$ ,  $f_y = x^2 + 2xy - x$ . If we set both partial derivatives equal

$$f_{xy}(x, y) = \lim_{t \rightarrow 0} \frac{1}{t^2} \left( (f(x + t, y + t) - f(x, y + t)) - (f(x + t, y) - f(x, y)) \right),$$

if the limit on the right hand side exists (notice we cannot take this as granted without further arguments).

Consider the expression from which the last limit is taken as the function  $\varphi(x, y, t)$ , and try to express it in terms of partial derivatives. For a temporarily fixed  $t$ , denote

$$g(x, y) = f(x + t, y) - f(x, y).$$

Notice that the partial derivative  $g_y$  exists on a neighborhood of  $(x, y)$  and it is continuous in  $(x, y)$ . Thus, for small  $t$ , we may apply the mean value theorem viewing  $g$  as function of  $y$ . Therefore, the expression in the last large parentheses equals to

$$g(x, y + t) - g(x, y) = t \cdot g_y(x, y + t_0)$$

for a suitable  $t_0$  which lies between 0 and  $t$  (the value of  $t_0$  depends on  $t$ ). Next,  $g_y(x, y) = f_y(x + t, y) - f_y(x, y)$ , so we may rewrite  $\varphi$  as

$$\begin{aligned} \varphi(x, y, t) &= \frac{1}{t} g_y(x, y + t_0) \\ &= \frac{1}{t} (f_y(x + t, y + t_0) - f_y(x, y + t_0)). \end{aligned}$$

Another application of the mean value theorem yields

$$\varphi(x, y, t) = f_{yx}(x + t_1, y + t_0)$$

for a suitable  $t_1$  between 0 and  $t$ . Since it is assumed that the second order partial derivatives of  $f$  are continuous at the point  $(x, y)$ , the requested limit  $\lim_{t \rightarrow 0} \varphi(x, y, t)$  must exist and we have arrived at the desired equality  $f_{xy} = f_{yx}$ , too.

**8.1.12. Schwarz's Theorem.** The same procedure for functions of  $n$  variables proves the following fundamental result:<sup>2</sup>

#### COMMUTATIVITY OF PARTIAL DERIVATIVES

**Theorem.** Let  $f : E_n \rightarrow \mathbb{R}$  be a  $k$ -times differentiable function with continuous partial derivatives up to order  $k$  (inclusive) at the point  $x \in \mathbb{R}^n$ . Then all partial derivatives of the function  $f$  at the point  $x$  up to order  $k$  (inclusive) are independent of the order of differentiation.



**PROOF.** The proof for the second order is illustrated above in the special case when  $n = 2$ . In fact, it yields the general case as well. Indeed, notice that for every fixed choice of a pair of coordinates  $x_i$  and  $x_j$ , the discussion of their interchanging takes

<sup>2</sup>This is a great example of a result which was used widely for many decades before a complete proof was published in 1873 by Karl Hermann Amandus Schwarz (1843-1921). The result is also known as Clairaut's theorem – following the earlier version for functions requesting continuous partial derivatives on a neighborhood of  $x$ .



to zero simultaneously, the system has the following solution:  $\{x = y = 0\}$ ,  $\{x = 0, y = 1\}$ ,  $\{x = 1, y = 0\}$ ,  $\{x = 1/3, y = 1/3\}$ , which are four stationary points of the given function.

The Hessian of the function  $f$  is

$$\begin{pmatrix} 2y & 2x + 2y - 1 \\ 2x + 2y - 1 & 2x \end{pmatrix}.$$

Its values at the stationary points are, respectively,

$$\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 2/3 & 1 \\ 1 & 2/3 \end{pmatrix}.$$

Therefore, the first three Hessians are indefinite, and the last one is positive definite. The point  $[1/3, 1/3]$  is thus a local minimum.  $\square$

**8.F.2.** Determine the point in the plane  $x+y+3z = 5$  lying in  $\mathbb{R}^3$  which is closest to the origin of the coordinate system. First, do this by applying the methods of linear algebra; then, using the methods of differential calculus.

**Solution.** It is the intersection point of the perpendicular going through the point  $[0, 0, 0]$  to the plane. The normal to the plane is  $(t, t, 3t)$ ,  $t \in \mathbb{R}$ . Substituting into the equation of the plane, we get the intersection point  $[5/11, 5/11, 15/11]$ .

Alternatively, we can minimize the distance (or its square) of the plane's points from the origin, i. e., the function

$$(5 - y - 3z)^2 + y^2 + z^2.$$

Setting the partial derivatives equal to zero, we get the system

$$3y + 10z - 15 = 0$$

$$2y + 3z - 5 = 0,$$

whose solution is as above. Since we know that the minimum exists and is the only stationary point, we need not calculate the Hessian any more.  $\square$

**8.F.3.** Determine the local extrema of the function

$$f(x, y) = x^2 + \arctan^2 x + |y^3 + y|, \quad x, y \in \mathbb{R}.$$

**Solution.** The function  $f$  can be written as the sum  $f_1 + f_2$ , where  $f_1(x) = x^2 + \arctan^2 x$ ,  $f_2(y) = |y^3 + y|$ ,  $x, y \in \mathbb{R}$ .

If the function  $f$  has a local extremum at a point, then it does so with respect to an arbitrary subset of its domain. In other words, if the function has, for instance, a maximum at a point  $[a, b]$  and we set  $y = b$ , then the univariate function  $f(x, b)$  of  $x$  must have a maximum at the point  $x = a$ . Let us thus fix an arbitrary  $y \in \mathbb{R}$ . For this fixed value of  $y$ , we get a univariate function, which is a shift of the function  $f_1$ . This means that its maxima and minima are at the same points. However, it is easy to find the extrema of the function  $f_1$ . We can just realize that this function is even (it is the sum of two even functions, and the function  $y = \arctan^2 x$  is the product of two odd functions) and increasing for  $x \geq 0$  (the composition as well as the sum of increasing functions is again an increasing function). Therefore, it has a unique extremum, and that is a minimum at the point  $x = 0$ . Similarly, for any fixed value of  $x$ ,  $f$  is a shift of the function  $f_2$ , and  $f_2$  has a minimum at the point  $y = 0$ , which is its only extremum. We

place in a two-dimensional affine subspace, (all the other variables are considered to be constant and do not affect in the discussion). So neighbouring partial derivatives may be interchanged. This solves the problem in order two.

In the case of higher-order derivatives, the proof can be completed by induction on the order. Every order of the indices  $i_1, \dots, i_k$  can be obtained from a fixed one by several interchanges of adjacent pairs of indices.  $\square$

**8.1.13. Hessian.** The differential was introduced as the linear form  $df(x)$  which approximates the function  $f$  at a point  $x$  in the best possible way.

Similarly, a quadratic approximation of a function  $f : E_n \rightarrow \mathbb{R}$  is possible.



### HESSIAN

**Definition.** If  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a twice differentiable function, the symmetric matrix of functions

$$Hf(x) = \left( \frac{\partial^2 f}{\partial x_i \partial x_j} (x) \right) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} (x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} (x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} (x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n} (x) \end{pmatrix}$$

is called the Hessian of the function  $f$ .

It is already seen from the previous reasonings that the vanishing of the differential at a point  $(x, y) \in E_2$  guarantees stationary behaviour along all curves going through this point. The Hessian

$$Hf(x, y) = \begin{pmatrix} f_{xx}(x, y) & f_{xy}(x, y) \\ f_{yx}(x, y) & f_{yy}(x, y) \end{pmatrix}$$

plays the role of the second derivative.

For every parametrized straight line

$$c(t) = (x(t), y(t)) = (x_0 + \xi t, y_0 + \eta t),$$

the derivative of the univariate function  $\alpha(t) = f(x(t), y(t))$  can be computed by means of the formula  $\frac{d}{dt} f(t) = f_x(x(t), y(t))x'(t) + f_y(x(t), y(t))y'(t)$  (derived in 8.1.8) and so the function

$$\begin{aligned} \beta(t) &= f(x_0, y_0) + t \frac{\partial f}{\partial x}(x_0, y_0)\xi + t \frac{\partial f}{\partial y}(x_0, y_0)\eta \\ &+ \frac{t^2}{2} \left( f_{xx}(x_0, y_0)\xi^2 + 2f_{xy}(x_0, y_0)\xi\eta + f_{yy}(x_0, y_0)\eta^2 \right) \end{aligned}$$

shares the same derivatives up to the second order (inclusive) at the point  $t = 0$  (calculate this on your own!). The function  $\beta$  can be written in terms of vectors as

$$\beta(t) = f(x_0, y_0) + df(x_0, y_0)(tv) + \frac{1}{2} Hf(x_0, y_0)(tv, tv),$$

where  $v = (\xi, \eta)$  is the increment given by the derivative of the curve  $c(t)$ , and the Hessian is used as a symmetric 2-form.

This is an expression which looks like Taylor's theorem for univariate functions, namely the quadratic approximation

have thus proved that  $f$  can have a local extremum only at the origin. Since

$$f(0, 0) = 0, \quad f(x, y) > 0, \quad [x, y] \in \mathbb{R}^2 \setminus \{[0, 0]\},$$

the function  $f$  has a strict local (even global) minimum at the point  $[0, 0]$ .  $\square$

**8.F.4.** Examine the local extrema of the function

$$f(x, y) = (x + y^2) e^{\frac{x}{2}}, \quad x, y \in \mathbb{R}.$$

**Solution.** This function has partial derivatives of all orders on the whole of its domain. Therefore, local extrema can occur only at stationary points, where both the partial derivatives  $f_x, f_y$  are zero. Then, it can be determined whether the local extremum occurs by computing the second derivatives.

We can easily determine that

$$\begin{aligned} f_x(x, y) &= e^{\frac{x}{2}} + \frac{1}{2}(x + y^2) e^{\frac{x}{2}}, \\ f_y(x, y) &= 2y e^{\frac{x}{2}}, \quad x, y \in \mathbb{R}. \end{aligned}$$

A stationary point  $[x, y]$  must satisfy

$$f_y(x, y) = 0, \quad \text{i. e.} \quad y = 0,$$

and, further,

$$f_x(x, y) = f_x(x, 0) = e^{\frac{x}{2}} \left(1 + \frac{1}{2}x\right) = 0, \quad \text{i. e.} \quad x = -2.$$

We can see that there is a unique stationary point, namely  $[-2, 0]$ .

Now, we calculate the Hessian  $Hf$  at this point. If this matrix (the corresponding quadratic form) is positive definite, the extremum is a strict local minimum. If it is negative definite, the extremum is a strict local maximum. Finally, if the matrix is indefinite, there will be no extremum at the point. We have

$$\begin{aligned} f_{xx}(x, y) &= \frac{1}{2} e^{\frac{x}{2}} \left(2 + \frac{1}{2}(x + y^2)\right), \quad f_{yy}(x, y) = 2 e^{\frac{x}{2}}, \\ f_{xy}(x, y) &= f_{yx}(x, y) = y e^{\frac{x}{2}}, \quad x, y \in \mathbb{R}. \end{aligned}$$

Therefore,

$$\begin{aligned} Hf(-2, 0) &= \begin{pmatrix} f_{xx}(-2, 0) & f_{xy}(-2, 0) \\ f_{yx}(-2, 0) & f_{yy}(-2, 0) \end{pmatrix} \\ &= \begin{pmatrix} 1/2e & 0 \\ 0 & 2/e \end{pmatrix}. \end{aligned}$$

We should recall that the eigenvalues of a diagonal matrix are exactly the values on the diagonal. Further, positive definiteness means that all the eigenvalues are positive. Hence it follows that there is a strict local minimum at the point  $[-2, 0]$ .  $\square$

**8.F.5.** Find the local extrema of the function

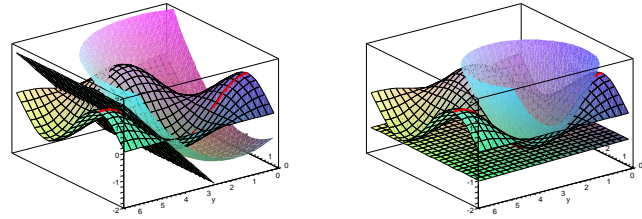
$$f(x, y, z) = x^3 + y^2 + \frac{z^2}{2} - 3xz - 2y + 2z, \quad x, y, z \in \mathbb{R}.$$

**Solution.** The function  $f$  is a polynomial; therefore, it has partial derivatives of all orders. It thus suffices to look for its stationary points (the extrema cannot be elsewhere). In order to find them, we differentiate  $f$  with respect to each of the three variables  $x, y, z$  and set the derivatives equal to zero. We thus obtain

$$3x^2 - 3z = 0, \quad \text{i. e.,} \quad z = x^2,$$

of a function by Taylor's polynomial of degree two. The following illustration shows both the tangent plane and this quadratic approximation for two distinct points and the function  $f(x, y) = \sin(x) \cos(y)$ .

popis obrázku



**8.1.14. Taylor's expansion.** The multidimensional version



of Taylor's theorem is an example of a mathematical statement where the most difficult part is finding the right formulation. The proof is then quite simple.

The discussion on the Hessians continues. Write  $D^k f$  for the  $k$ -th order approximations of the function  $f : E_n \rightarrow \mathbb{R}^n$ . It is always a  $k$ -linear expressions in the increments.

The differential  $D^1 f = df$  (the first order) and the Hessian  $D^2 f = Hf$  (the second order) are already discussed. For functions  $f : E_n \rightarrow \mathbb{R}$ , points  $x = (x_1, \dots, x_n) \in E_n$ , and increments  $v = (\xi_1, \dots, \xi_n)$ , set

$$D^k f(x)(v) = \sum_{1 \leq i_1, \dots, i_k \leq n} \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x_1, \dots, x_n) \xi_{i_1} \cdots \xi_{i_k}.$$

An illustrative example (making use of the symmetry of the partial derivatives) is, for  $E_2$ , the third-order expression

$$\begin{aligned} D^3 f(x, y)(\xi, \eta) &= \frac{\partial^3 f}{\partial x^3} \xi^3 + 3 \frac{\partial^3 f}{\partial x^2 \partial y} \xi^2 \eta \\ &\quad + 3 \frac{\partial^3 f}{\partial x \partial y^2} \xi \eta^2 + \frac{\partial^3 f}{\partial y^3} \eta^3, \end{aligned}$$

and, in general,

$$D^k f(x, y)(\xi, \eta) = \sum_{\ell=0}^k \binom{k}{\ell} \frac{\partial^k f}{\partial x^{k-\ell} \partial y^\ell} \xi^{k-\ell} \eta^\ell.$$

#### TAYLOR'S EXPANSION WITH REMAINDER

**Theorem.** Let  $f : E_n \rightarrow \mathbb{R}$  be a  $k$ -times differentiable function in a neighbourhood  $\mathcal{O}_\delta(x)$  of a point  $x \in E_n$ . For every increment  $v \in \mathbb{R}^n$  of size  $\|v\| < \delta$ , there exists a number  $\theta, 0 \leq \theta \leq 1$ , such that

$$\begin{aligned} f(x + v) &= f(x) + D^1 f(x)(v) + \frac{1}{2!} D^2 f(x)(v) + \\ &\quad \cdots + \frac{1}{(k-1)!} D^{k-1} f(x)(v) + \frac{1}{k!} D^k f(x + \theta \cdot v)(v). \end{aligned}$$

$$2y - 2 = 0, \quad \text{i. e., } y = 1,$$

and (utilizing the first equation)

$$z - 3x + 2 = 0, \quad \text{i. e., } x \in \{1, 2\}.$$

Therefore, there are two stationary points, namely  $[1, 1, 1]$  and  $[2, 1, 4]$ . Now, we compute all second-order partial derivatives:

$$\begin{aligned} f_{xx} &= 6x, & f_{xy} &= f_{yx} = 0, & f_{xz} &= f_{zx} = -3, \\ f_{yy} &= 2, & f_{yz} &= f_{zy} = 0, & f_{zz} &= 1. \end{aligned}$$

Having this, we are able to evaluate the Hessian at the stationary points:

$$\begin{aligned} Hf(1, 1, 1) &= \begin{pmatrix} 6 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{pmatrix}, \\ Hf(2, 1, 4) &= \begin{pmatrix} 12 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{pmatrix}. \end{aligned}$$

Now, we need to know whether these matrices are positive definite, negative definite, or indefinite in order to determine whether and which extrema occur at the corresponding points. Clearly, the former matrix (for the point  $[1, 1, 1]$ ) has eigenvalue  $\lambda = 2$ . Since its determinant equals  $-6$  and it is a symmetric matrix (all eigenvalues are real), the matrix must have a negative eigenvalue as well (because the determinant is the product of the eigenvalues). Therefore, the matrix  $Hf(1, 1, 1)$  is indefinite, and there is no extremum at the point  $[1, 1, 1]$ .

We will use the so-called Sylvester's criterion for the latter matrix  $Hf(2, 1, 4)$ . According to this criterion, a real-valued symmetric matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{12} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{13} & a_{23} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & a_{3n} & \cdots & a_{nn} \end{pmatrix}$$

is positive definite if and only if all of its leading principal minors  $A$ , i. e. the determinants

$$d_1 = |a_{11}|, \quad d_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}, \quad d_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix}, \quad \dots$$

$$d_n = |A|,$$

are positive. Further, it is negative definite iff

$$d_1 < 0, \quad d_2 > 0, \quad d_3 < 0, \quad \dots, \quad (-1)^n d_n > 0.$$

The inequalities

$$|12| = 12 > 0, \quad \begin{vmatrix} 12 & 0 \\ 0 & 2 \end{vmatrix} = 24 > 0, \quad \begin{vmatrix} 12 & 0 & -3 \\ 0 & 2 & 0 \\ -3 & 0 & 1 \end{vmatrix} = 6 > 0,$$

imply that the matrix  $Hf(2, 1, 4)$  is positive definite – there is a strict local minimum at the point  $[2, 1, 4]$ .  $\square$

**8.F.6.** Find the local extrema of the function

$$z = (x^2 - 1)(1 - x^4 - y^2), \quad x, y \in \mathbb{R}.$$

**PROOF.** Given an increment  $v \in \mathbb{R}^n$ , consider the parametrized straight line  $c(t) = x + tv$  in  $E_n$ , and examine the function  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  defined by the composition  $\varphi(t) = f \circ c(t)$ . Taylor's theorem for univariate functions claims that (see Theorem 6.1.3)



$$\begin{aligned} \varphi(t) &= \varphi(0) + \varphi'(0)t + \dots \\ &\quad + \frac{1}{(k-1)!} \varphi^{(k-1)}(0)t^{k-1} + \frac{1}{k!} \varphi^{(k)}(\theta)t^k. \end{aligned}$$

It remains to verify that computing the derivatives  $\varphi^{(\ell)}$  yields the desired relation. This can be done quite easily by induction on the order  $k$ .

For  $k = 1$ , Taylor's theorem coincides with the corollary of the mean value theorem applied to the directional derivative, which is already used several times. When deriving it, the formula

$$\frac{d}{dt} \varphi(t) = \frac{\partial f}{\partial x_1}(x(t)) \cdot x'_1(t) + \dots + \frac{\partial f}{\partial x_n}(x(t)) \cdot x'_n(t)$$

is used. It holds for every continuously differentiable curve and function  $f$ , cf. 8.1.8 and 8.1.9. This means that

$$\varphi'(t) = D^1 f(c(t))(c'(t)) = D^1 f(c(t))(v)$$

for all  $t$  in a neighbourhood of zero. Proceed similarly for functions  $D^\ell f$ . Write  $c'(t)$  instead of the increment  $v$ , and recall that further differentiation of  $c(t)$  leads identically to zero everywhere, i.e.  $c''(t) = 0$  for all  $t$  (since it is a parametrized straight line). Suppose

$$\begin{aligned} \varphi^{(\ell)}(t) &= D^\ell f(x(t))(v) \\ &= \sum_{i_1, \dots, i_\ell} \left( \frac{\partial^\ell f}{\partial x_{i_1} \dots \partial x_{i_\ell}}(x_1(t), \dots, x_n(t)) x'_{i_1}(t) \cdots x'_{i_\ell}(t) \right) \end{aligned}$$

and calculate  $\varphi^{(\ell+1)}(t)$ . By the above formula for first-order differentiation in a given direction and the rule for the derivative of a product (see Theorem 5.3.4), the differentiation of the composite function gives

$$\begin{aligned} \varphi^{(\ell+1)}(t) &= \frac{d}{dt} D^\ell f(c(t))(c'(t)) \\ &= \frac{d}{dt} \sum_{i_1, \dots, i_\ell} \left( \frac{\partial^\ell f}{\partial x_{i_1} \dots \partial x_{i_\ell}}(x_1(t), \dots, x_n(t)) \right. \\ &\quad \left. \cdot x'_{i_1}(t) \cdots x'_{i_\ell}(t) \right) \\ &= \sum_{i_1, \dots, i_\ell} \left( \sum_{j=1}^n \frac{\partial^{\ell+1} f}{\partial x_{i_1} \dots \partial x_{i_\ell} \partial x_j}(x_1(t), \dots, x_n(t)) \right. \\ &\quad \left. \cdot x'_j(t) \cdot x'_{i_1}(t) \cdots x'_{i_\ell}(t) \right) + 0, \end{aligned}$$

which is the required formula for order  $\ell+1$ . Taylor's theorem now follows from substituting into the equality for  $\varphi$  at the beginning of this proof and the enumeration at the right values of  $t$ .  $\square$

**Solution.** Once again, we calculate the partial derivatives  $z_x$ ,  $z_y$  and set them equal to zero. This leads to the equations

$$-6x^5 + 4x^3 + 2x - 2xy^2 = 0, \quad (x^2 - 1)(-2y) = 0,$$

whose solutions  $[x, y] = [0, 0]$ ,  $[x, y] = [1, 0]$ ,  $[x, y] = [-1, 0]$ . (In order to find these solutions, it suffices to find the real roots  $1, -1$  of the polynomial  $-6x^4 + 4x^2 + 2$  using the substitution  $u = x^2$ . Now, we compute the second-order partial derivatives

$$\begin{aligned} z_{xx} &= -30x^4 + 12x^2 + 2 - 2y^2, \\ z_{xy} &= z_{yx} = -4xy, \\ z_{yy} &= -2(x^2 - 1) \end{aligned}$$

and evaluate the Hessian at the stationary points:

$$\begin{aligned} Hz(0, 0) &= \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \\ Hz(1, 0) &= Hz(-1, 0) = \begin{pmatrix} -16 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

We can see that the first matrix is positive definite, so the function has a strict local minimum at the origin.

However, the second and third matrices are negative semidefinite. Therefore, the knowledge of second partial derivatives is insufficient for deciding whether there is an extremum at the points  $[1, 0]$  and  $[-1, 0]$ . On the other hand, we can examine the function values near these points. We have  $z(1, 0) = z(-1, 0) = 0$ ,  $z(x, 0) < 0$  for  $x \in (-1, 1)$ . Further, consider  $y$  dependent on  $x \in (-1, 1)$  by the formula  $y = \sqrt{2(1 - x^4)}$ , so that  $y \rightarrow 0$  for  $x \rightarrow \pm 1$ . For this choice, we get  $z(x, \sqrt{2(1 - x^4)}) = (x^2 - 1)(x^4 - 1) > 0$ ,  $x \in (-1, 1)$ . We have thus shown that in arbitrarily small neighborhoods of the points  $[1, 0]$  and  $[-1, 0]$ , the function  $z$  takes on both higher and lower values than the function value at the corresponding point. Therefore, these are not extrema.  $\square$

**8.F.7.** Decide whether the polynomial

$$p(x, y) = x^6 + y^8 + y^4x^4 - x^6y^5$$

has a local extremum at the stationary point  $[0, 0]$ .

**Solution.** We can easily verify that the partial derivatives  $p_x$  and  $p_y$  are indeed zero at the origin. However, each of the partial derivatives  $p_{xx}$ ,  $p_{xy}$ ,  $p_{yy}$  is also equal to zero at the point  $[0, 0]$ . The Hessian  $Hp(0, 0)$  is thus both positive and negative semidefinite at the same time. However, a simple idea can lead us to the result: We can notice that  $p(0, 0) = 0$  and

$$p(x, y) = x^6(1 - y^5) + y^8 + y^4x^4 > 0$$

for  $[x, y] \in \mathbb{R} \times (-1, 1) \setminus \{[0, 0]\}$ . Therefore, the given polynomial has a local minimum at the origin.  $\square$

**8.F.8.** Determine local extrema of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^2y + y^2z + x - z$  on  $\mathbb{R}^3$ .  $\circ$

**8.F.9.**

**8.1.15. Formula with multi-indices.** To simplify the notation, let us introduce the multi-index notation for the polynomials with more variables.

MULTI-INDICES

A *multi-index*  $\alpha$  of length  $n$  is an  $n$ -tuple of non-negative integers  $(\alpha_1, \dots, \alpha_n)$ . The integer  $|\alpha| = \alpha_1 + \dots + \alpha_n$  is called the *size* of the multi-index  $\alpha$ .

Monomials are written shortly as  $x^\alpha$  instead of  $x_1^{\alpha_1}x_2^{\alpha_2}\dots x_n^{\alpha_n}$ . Real polynomials in  $n$  variables can be symbolically expressed in a similar way as univariate polynomials:

$$f = \sum_{|\alpha| \leq k} a_\alpha x^\alpha, \quad g = \sum_{|\beta| \leq \ell} b_\beta x^\beta \in \mathbb{R}[x_1, \dots, x_n].$$

$f$  is said to have total degree  $k$  if at least one coefficient with multi-indices  $\alpha$  of size  $k$  is non-zero, while all the coefficients with multi-indices of larger sizes vanish.

Nice formulae express addition and multiplication of multivariate polynomials of degrees  $k$  and  $\ell$  respectively:

$$\begin{aligned} f + g &= \sum_{|\alpha| \leq \max(k, \ell)} (a_\alpha + b_\alpha)x^\alpha, \\ fg &= \sum_{|\gamma|=0}^{k+\ell} \left( \sum_{\alpha+\beta=\gamma} a_\alpha b_\beta \right) x^\gamma, \end{aligned}$$

where the multi-indices are added componentwise, and the formally non-existing coefficients are assumed to be zero.

Moreover we write shortly

$$\partial_\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}$$

and  $\alpha! = \alpha_1! \dots \alpha_n!$ . In particular,  $\partial_\alpha f = f$  if  $|\alpha| = 0$ .

TAYLOR'S POLYNOMIALS VIA MULTI-INDICES

Taylor's expansion up to order  $r$  of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , for an increment  $v \in \mathbb{R}^n$  is the polynomial

$$(1) \quad f(x + v) = f(x) + \sum_{1 \leq |\alpha| \leq k} \frac{1}{\alpha!} \partial_\alpha f(x) v^\alpha,$$

quite as the formula in dimension one.

**8.1.16. Real analytic functions.** If the multidimensional power series  $F(v) = \sum_{|\alpha| \geq 0} \frac{1}{\alpha!} \partial_\alpha f(x) v^\alpha$  converges at some neighborhood of  $v = 0$ , we call the function  $f$  (real) *analytic* on a neighborhood of  $x$ .

For instance, this happens if all the partial derivatives are uniformly bounded, i.e.,  $\partial_\alpha f(x) < C$  for all  $\alpha$ . Indeed, we may estimate

$$\left| \frac{1}{k!} \sum_{i_1, \dots, i_k} \frac{\partial^k f(x)}{\partial x_{i_1} \dots \partial x_{i_k}} v_{i_1} \dots v_{i_k} \right| \leq \frac{1}{k!} n^k C \|v\|^k$$

and thus the Taylor's series converges absolutely by the Weierstrass criterion for all  $v$ . Think about the details. Actually, our argument shows that the Taylor's series converges (on a small

Determine the local extrema of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^2y - y^2z + 4x + z$  on  $\mathbb{R}^3$ . ○

**8.F.10.** Determine the local extrema of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = xz^2 + y^2z - x + y$  on  $\mathbb{R}^3$ . ○

**8.F.11.** Determine the local extrema of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = y^2z - xz^2 + x + 4y$  on  $\mathbb{R}^3$ . ○

**8.F.12.** Determine the local extrema of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2y + x^2 + 2y^2 + y$  on  $\mathbb{R}^2$ . ○

**8.F.13.** Determine the local extrema of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2y + 2y^2 + 2y$  on  $\mathbb{R}^2$ . ○

**8.F.14.** Determine the local extrema of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2 + xy + 2y^2 + y$  on  $\mathbb{R}^2$ . ○

**8.F.15.** Determine the local extrema of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^2 + xy - 2y^2 + y$  on  $\mathbb{R}^2$ . ○

**G. Implicitly given functions and mappings**

**8.G.1.** Let  $F : \mathbb{R}^2 \rightarrow \mathbb{R}$  be a function,  $F(x, y) = xy \sin(\frac{\pi}{2}xy^2)$ . Show that the equality  $F(x, y) = 1$  implicitly defines a function  $f : U \rightarrow \mathbb{R}$  on a neighborhood  $U$  of the point  $[1, 1]$  so that  $F(x, f(x)) = 1$  for  $x \in U$ . Determine  $f'(1)$ .

**Solution.** The function is differentiable on the whole  $\mathbb{R}^2$ , so it is such on any neighborhood of the point  $[1, 1]$ . Let us evaluate  $F_y$  at  $[1, 1]$ :

$$F_y(x, y) = x \sin\left(\frac{\pi}{2}xy^2\right) + \pi x^2 y^2 \cos\left(\frac{\pi}{2}xy^2\right),$$

so  $F_y(1, 1) = 1 \neq 0$ . Therefore, it follows from theorem 8.1.25 that the equation  $F(x, y) = 1$  implicitly determines on a neighborhood of the point  $(1, 1)$  a function  $f : U \rightarrow \mathbb{R}$  defined on a neighborhood of the point (number) 1. Moreover, we have

$$F_x(x, y) = y \sin\left(\frac{\pi}{2}xy^2\right) + \frac{\pi}{2}xy^3 \cos\left(\frac{\pi}{2}xy^2\right),$$

so the derivative of the function  $f$  at the point 1 satisfies

$$f'(1) = -\frac{F_x(1, 1)}{F_y(1, 1)} = -\frac{1}{1} = -1. \quad \square$$

**Remark.** Notice that although we are unable to explicitly define the function  $f$  from the equation  $F(x, f(x)) = 1$ , we are able to determine its derivative at the point 1.

**8.G.2.** Considering the function  $F : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,

$$F(x, y) = e^x \sin(y) + y - \pi/2 - 1,$$

show that the equation  $F(x, y) = 0$  implicitly defines the variable  $y$  to be a function of  $x$ ,  $y = f(x)$ , on a neighborhood of the point  $[0, \pi/2]$ . Compute  $f'(0)$ .

**Solution.** The function is differentiable in a neighborhood of the point  $[0, \pi/2]$ ; moreover,  $F_y = e^x \cos y + 1$ ,  $F(0, \pi/2) = 1 \neq 0$ , so the equation indeed defines a function  $f : U \rightarrow \mathbb{R}$  on a neighborhood of the point  $[0, \pi/2]$ . Further, we have  $F_x = e^x \sin y$ ,  $F_x(0, \pi/2) = 1$ , and its derivative at the point

neighborhood of  $x$ ) if the partial derivatives grow with the order  $k$  slower than  $k!$ .

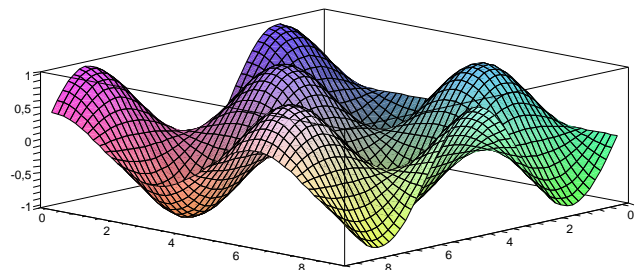
**8.1.17. Local extrema.** We examine the local maxima and minima of functions on  $E_n$  using the differential and the Hessian. Just as in the case of univariate functions, an interior point  $x_0 \in E_n$  of the domain of a function  $f$  is said to be a (local) *maximum* or *minimum* if and only if there is a neighbourhood  $U$  of  $x_0$  such that for all points  $x \in U$ , the function value satisfies  $f(x) \leq f(x_0)$  or  $f(x) \geq f(x_0)$ , respectively. If strict inequalities hold for all  $x \neq x_0$ , there is a *strict extremum*.

To simplify, suppose that  $f$  has continuous both first-order and second-order partial derivatives on its domain. A necessary condition for the existence of an extremum at a point  $x_0$  is that the differential be zero at this point, i.e.,  $df(x_0) = 0$ . If  $df(x_0) \neq 0$ , then there is a direction  $v$  in which  $d_v f(x_0) \neq 0$ . However, then the function value is increasing at one side of the point  $x_0$  along the line  $x_0 + tv$  and it is decreasing on the other side, see 5.3.2.

An interior point  $x \in E_n$  of the domain of a function  $f$  at which the differential  $df(x)$  is zero is called a *stationary point of the function*  $f$ .

To illustrate the concept on a simple function in  $E_2$ , consider  $f(x, y) = \sin(x) \cos(y)$ .

The shape of this function resembles the well-known egg plates, so it is evident that there are many extrema, and also many stationary points which are not extrema (“saddles” are visible in the picture).



Calculate the first derivatives, and then the necessary second-order ones:

$$f_x(x, y) = \cos(x) \cos(y), \quad f_y(x, y) = -\sin(x) \sin(y),$$

and both derivatives are zero for two sets of points

- (1)  $\cos(x) = 0, \sin(y) = 0$ , that is  $(x, y) = (\frac{2k+1}{2}\pi, \ell\pi)$ , for any  $k, \ell \in \mathbb{Z}$
- (2)  $\cos(y) = 0, \sin(x) = 0$ , that is  $(x, y) = (k\pi, \frac{2\ell+1}{2}\pi)$ , for any  $k, \ell \in \mathbb{Z}$ .

The second partial derivatives are

$$Hf(x, y) = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{pmatrix} (x, y) = \begin{pmatrix} -\sin(x) \cos(y) & -\cos(x) \sin(y) \\ -\cos(x) \sin(y) & -\sin(x) \cos(y) \end{pmatrix}.$$

So the following Hessians are obtained in two sets of stationary points:

0 satisfies:

$$f'(0) = -\frac{F_x(0, \pi/2)}{F_y(0, \pi/2)} = -\frac{1}{1} = -1. \quad \square$$

**8.G.3.** Let

$$F(x, y, z) = \sin(xy) + \sin(yz) + \sin(xz).$$

Show that the equation  $F(x, y, z) = 0$  implicitly defines a function  $z(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$  on a neighborhood of the point  $[\pi, 1, 0] \in \mathbb{R}^3$  so that  $F(x, y, z(x, y)) = 0$ .

Determine  $z_x(\pi, 1)$  and  $z_y(\pi, 1)$ .

**Solution.** We will calculate  $F_z = y \cos(yz) + x \cos(xz)$ ,  $F_z(\pi, 1, 0) = \pi + 1 \neq 0$ , and the function  $z(x, y)$  is defined by the equation  $F(x, y, z(x, y)) = 0$  on a neighborhood of the point  $[\pi, 1, 0]$ . In order to find the values of the wanted partial derivatives, we first need to calculate the values of the remaining partial derivatives of the function  $F$  at the point  $[\pi, 1, 0]$ .

$$F_x(x, y, z) = y \cos(xy) + z \cos(xz) \quad F_x(\pi, 1, 0) = -1,$$

$$F_y(x, y, z) = x \cos(xy) + z \cos(yz) \quad F_y(\pi, 1, 0) = -\pi,$$

odkud

$$z_x(\pi, 1) = -\frac{F_x(\pi, 1, 0)}{F_z(\pi, 1, 0)} = -\frac{1}{\pi + 1},$$

$$z_y(\pi, 1) = -\frac{F_y(\pi, 1, 0)}{F_z(\pi, 1, 0)} = -\frac{\pi}{\pi + 1}. \quad \square$$

**8.G.4.** Having the mapping  $F : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ ,  $F(x, y, z) = (f(x, y, z), g(x, y, z)) = (e^x \sin y, xyz)$ , show that the equation  $F(x, c_1(x), c_2(x)) = (0, 0)$  defines a curve  $c : \mathbb{R} \rightarrow \mathbb{R}^2$  on a neighborhood of the point  $[1, \pi, 1]$ . Determine the tangent vector to this curve at the point 1.

**Solution.** We will calculate the square matrix of the partial derivatives of the mapping  $F$  with respect to  $y$  and  $z$ :

$$H(x, y, z) = \begin{pmatrix} f_y & f_z \\ g_y & g_z \end{pmatrix} = \begin{pmatrix} x \cos y e^x \sin y & 0 \\ xz & xy \end{pmatrix}.$$

Hence,  $H(1, \pi, 1) = \begin{pmatrix} -1 & 0 \\ 1 & \pi \end{pmatrix}$  and  $\det H(1, \pi, 1) = -\pi \neq 0$ . Now, it follows from the implicit mapping theorem (see 8.1.25) that the equation  $F(x, c_1(x), c_2(x)) = (0, 0)$  on a neighborhood of the point  $[1, \pi, 1]$  determines a curve  $(c_1(x), c_2(x))$  defined on a neighborhood of the point  $[1, \pi]$ . In order to find its tangent vector at this point, we need to determine the (column) vector  $(f_x, g_x)$  at this point:

$$\begin{pmatrix} f_x \\ g_x \end{pmatrix} = \begin{pmatrix} \sin y e^x \sin y \\ yz \end{pmatrix}, \quad \begin{pmatrix} f_x(1, \pi, 1) \\ g_x(1, \pi, 1) \end{pmatrix} = \begin{pmatrix} 0 \\ \pi \end{pmatrix}.$$

The wanted tangent vector is thus

$$\begin{aligned} \begin{pmatrix} (c_1)_x(1) \\ (c_2)_x(1) \end{pmatrix} &= \begin{pmatrix} f_y(1, \pi, 1) & f_z(1, \pi, 1) \\ g_y(1, \pi, 1) & g_z(1, \pi, 1) \end{pmatrix}^{-1} \begin{pmatrix} f_x(1, \pi, 1) \\ g_x(1, \pi, 1) \end{pmatrix} = \\ &= \begin{pmatrix} -1 & 0 \\ 1 & \pi \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \pi \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ \frac{1}{\pi} & \frac{1}{\pi} \end{pmatrix} \begin{pmatrix} 0 \\ \pi \end{pmatrix}. \end{aligned}$$

(1)  $Hf(k\pi + \frac{\pi}{2}, \ell\pi) = \pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ , where the minus sign occurs when  $k$  and  $\ell$  have the same parity (remainder on division by two), and the sign  $+$  occurs in the other case;

(2)  $Hf(k\pi, \ell\pi + \frac{\pi}{2}) = \pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ , where again the minus sign occurs when  $k$  and  $\ell$  have the same parity, and the sign  $+$  occurs in the other case;

From the proposition of Taylor's theorem for order  $k = 2$ , there is, in a neighbourhood of one of the stationary points  $(x_0, y_0)$ ,

$$f(x, y) = f(x_0, y_0) + \frac{1}{2}Hf(x_0 + \theta(x - x_0), y_0 + \theta(y - y_0))(\xi, \eta).$$

Here,  $Hf$  is considered to be a quadratic form evaluated at the increment  $(x - x_0, y - y_0) = (\xi, \eta)$ . In the case (1),  $Hf(x_0, y_0)(\xi, \eta) = \pm(\xi^2 + \eta^2)$ , while in the case (2),  $Hf(x_0, y_0)(\xi, \eta) = \pm 2\xi\eta$ . While in the first case, the quadratic form is either always positive or always negative on all nonzero arguments, in the second case, there are always arguments with positive values and other arguments with negative values.

Since the Hessian of the function is continuous (i.e. all the partial derivatives up to order two are continuous), the Hessians in the nearby points are small perturbations of those in  $(x_0, y_0)$  and so these properties of the quadratic form  $Hf(x, y)$  remain true on some neighbourhood of  $(x_0, y_0)$ . This is obvious in cases (1) and (2), since a small perturbation of the matrices does clearly not change the latter properties of the quadratic forms in question. A general formal proof is presented below.

The local maximum occurs if and only if the point  $(x_0, y_0)$  belongs to the case (1) with  $k$  and  $\ell$  of the same parity. On the other hand, if the parities are different, then the point from the case (1) happens to be a point of a local minimum.

On the other hand, in the case (2) the entire function  $f$  behaves similarly to the Hessian and so the "saddle" points are not extrema.

**8.1.18. The decision rules.** In order to formulate the general statement about the Hessian and the local extrema at stationary points, it is necessary to remember the discussion about quadratic forms from the paragraphs 4.3.2–4.3.3 in the chapter on affine geometry. There are introduced the following types of quadratic forms  $h : E_n \rightarrow \mathbb{R}$ :

- *positive definite* if and only if  $h(u) > 0$  for all  $u \neq 0$
- *positive semidefinite* if and only if  $h(u) \geq 0$  for all  $u \in V$
- *negative definite* if and only if  $h(u) < 0$  for all  $u \neq 0$
- *negative semidefinite* if and only if  $h(u) \leq 0$  for all  $u \in V$
- *indefinite* if and only if  $h(u) > 0$  and  $h(v) < 0$  for appropriate  $u, v \in V$ .



- There are methods to allow determining whether or not a given form has any of these properties.

The Taylor expansion with remainder immediately yields the following rules:

LOCAL EXTREMA

**H. Constrained optimization**

We will begin with a somewhat atypical optimization problem.

**8.H.1.** A betting office accepts bets on the outcome of a tennis match. Let the odds laid against player  $A$  winning be  $a : 1$  (i. e., if a bettor bets  $x$  dollars on the event that player  $A$  wins and this really happens, then the bettor wins  $ax$  dollars) and, similarly, let the odds laid against player  $B$  winning be  $b : 1$  (fees are neglected). What is the necessary and sufficient condition for (positive real) numbers  $a$  and  $b$  so that a bettor cannot guarantee any profit regardless the actual outcome of the match? (For instance, if the odds were laid  $1.5 : 1$  against the win of  $A$  and  $5 : 1$  against the win of  $B$ , then the bettor could bet 3 dollars on  $B$  winning and 7 dollars on  $A$  winning and profit from this bet in either case).

**Solution.** Let the bettor have  $P$  dollars. The bet amount can be divided to  $kP$  and  $(1 - k)P$  dollars, where  $k \in (0, 1)$ . The profit is then  $akP$  dollars (if player  $A$  wins) or  $b(1 - k)P$  dollars (if  $B$  does). The bettor is always guaranteed to win the lesser of these two amounts; the total profit (or loss) is obtained by subtracting the bet  $P$ , then. Since each of  $a, b, P$  is a positive real number, the function  $akP$  is increasing, and the function  $b(1 - k)P$  is decreasing with respect to  $k$ . For  $k = 0$ ,  $b(1 - k)P$  is greater; for  $k = 1$ ,  $akP$  is. The minimum of the two numbers  $akP$  and  $b(1 - k)P$  is thus maximal for a  $k \in (0, 1)$ , namely for the value  $k_0$  which satisfies  $ak_0P = b(1 - k_0)P$ , whence  $k_0 = \frac{b}{a+b}$ . Therefore, the betting office must choose  $a, b$  so that  $ak_0P = b(1 - k_0)P < P$ , which is equivalent to  $ak_0 < 1$ , i. e.,  $ab < a + b$ . □

We managed to solve this constrained optimization problem even without using the differential calculus. However, we will not be able to do so in the following problems.

**8.H.2.** Find the extremal values of the function

$$h(x, y, z) = x^3 + y^3 + z^3$$

on the unit sphere  $S$  in  $\mathbb{R}^3$  given by the equation

$$F(x, y, z) = x^2 + y^2 + z^2 - 1$$

as well as on the circle which is the intersection of this sphere with the plane

$$G(x, y, z) = x + y + z.$$

**Solution.** First, we will look for stationary points of the function  $h$  on the sphere  $S$ . Computing the corresponding gradients (for instance,  $\text{grad } h(x, y, z) = (3x^2, 3y^2, 3z^2)$ ), we get the system

$$\begin{aligned} 0 &= 3x^2 - 2\lambda x, \\ 0 &= 3y^2 - 2\lambda y, \\ 0 &= 3z^2 - 2\lambda z, \\ 0 &= x^2 + y^2 + z^2 - 1 \end{aligned}$$

**Theorem.** Let  $f : E_n \rightarrow \mathbb{R}$  be a twice continuously differentiable function and  $x \in E_n$  be a stationary point of the function  $f$ . Then

- (1)  $f$  has a strict local minimum at  $x$  if  $Hf(x)$  is positive definite,
- (2)  $f$  has a strict local maximum at  $x$  if  $Hf(x)$  is negative definite,
- (3)  $f$  does not have an extremum at  $x$  if  $Hf(x)$  is indefinite.

**PROOF.** The Taylor second-order expansion with remainder applied to our function  $f(x_1, \dots, x_n)$ , an arbitrary point  $x = (x_1, \dots, x_n)$ , and any increment  $v = (v_1, \dots, v_n)$ , such that all points  $x + \theta v, \theta \in [0, 1]$ , lie in the domain of the function  $f$ , says that

$$f(x + v) = f(x) + df(x)(v) + \frac{1}{2}Hf(x + \theta \cdot v)(v)$$

for an appropriate real number  $\theta, 0 \leq \theta \leq 1$ . Since it is supposed that the differential is zero, we obtain

$$f(x + v) = f(x) + \frac{1}{2}Hf(x + \theta \cdot v)(v).$$

By assumption, the quadratic form  $Hf(x)$  is continuously dependent on the point  $x$ , and the definiteness or indefiniteness of quadratic forms can be determined by the sign of the major subdeterminants of the matrix  $Hf$ , see Sylvester's criterion in paragraph 4.3.3. However, the determinant itself is a polynomial expression in the coefficients of the matrix, hence a continuous function. Therefore, the non-vanishing and signs of the examined determinants are the same in a sufficiently small neighbourhood of the point  $x$  as at the point  $x$  itself.

In particular, for a positive definite  $Hf(x)$ , it is guaranteed that at a stationary point  $x, f(x + v) > f(x)$  for sufficiently small  $v$ . So this is a sharp minimum of the function  $f$  at the point  $x$ . The case of negative definiteness is analogous. If  $Hf(x)$  is indefinite, then there are directions  $v, w$  in which  $f(x + v) > f(x)$  and  $f(x + w) < f(x)$ , so there is no extremum at the stationary point in question. □

The theorem yields no result if the Hessian of the function is degenerate, yet not indefinite at the point in question. The reason is the same as in the case of univariate functions. In these cases, there are directions in which both the first and second derivatives vanish, so at this level of approximation, it cannot be determined whether the function behaves like  $t^3$  or  $\pm t^4$  until higher-order derivatives in the necessary directions are calculated.

At the same time, even at those points where the differential is non-zero, the definiteness of the Hessian  $Hf(x)$  has

consisting of four equations in four variables. Before trying to solve this system, we can estimate how many local constrained extrema we should anticipate the function to have. Surely,  $h(P)$  is in absolute value equal to at most 1, and this happens at all intersection points of the coordinate axes with  $S$ . Therefore, we are likely to get 6 local extrema. Further, inside every eighth of the sphere given by the coordinate planes, there may or may not be another extremum. The particular quadrants can be easily parametrized, and the function  $h$  (considered a function of two parameters) can be analyzed by standard means (or we can have it drawn in Maple, for example).

Actually, solving the system (no matter whether algebraically or in Maple again) leads to a great deal of stationary points. Besides the six points we have already talked about (two of the coordinates equal to zero and the other to  $\pm 1$ ) and which have  $\lambda = \pm \frac{3}{2}$ , there are also the points

$$P_{\pm} = \pm \left( \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3} \right),$$

for example, where a local extremum indeed occurs.

If we restrict our interest to the points of the circle  $K$ , we must give another function  $G$  another free parameter  $\eta$  representing the gradient coefficient. This leads to the bigger system

$$\begin{aligned} 0 &= 3x^2 - 2\lambda x - \eta, \\ 0 &= 3y^2 - 2\lambda y - \eta, \\ 0 &= 3z^2 - 2\lambda z - \eta, \\ 0 &= x^2 + y^2 + z^2 - 1, \\ 0 &= x + y + z. \end{aligned}$$

However, since a circle is also a compact set,  $h$  must have both a global minimum and maximum on it. Further analysis is left to the reader.  $\square$

**8.H.3.** Determine whether the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^2y$  has any extrema on the surface  $2x^2 + 2y^2 + z^2 = 1$ . If so, find these extrema and determine their types.

**Solution.** Since we are interested in extrema of a continuous function on a compact set (ellipsoid) – it is both closed and bounded in  $\mathbb{R}^3$  – the given function must have both a minimum and maximum on it. Moreover, since the constraint is given by a continuously differentiable function and the examined function is differentiable, the extrema must occur at stationary points of the function in question on the given set. We can build the following system for the stationary points:

$$\begin{aligned} 2xy &= 4kx, \\ x^2 &= 4ky, \\ 0 &= 2kz. \end{aligned}$$

This system is satisfied by the points  $[\pm \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{6}}, 0]$  and  $[\pm \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{6}}, 0]$ . The function takes on only two values at these four stationary points. It follows from the above that the first and second stationary points are maxima of

similar consequences as the non-vanishing of the second derivative of a univariate function. For a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , the expression  $z(x + v) = f(x) + df(x)(v)$  defines the tangent hyperplane to the graph of the function  $f$  in the space  $\mathbb{R}^{n+1}$ . Taylor's theorem of order two with remainder, as used in the proof above, provides the expression  $f(x + v) = z(x + v) + \frac{1}{2}Hf(x + \theta v)(v)$ .

If the Hessian is positive definite, all the values of the function  $f$  lie above the values of the tangent hyperplane for arguments in a sufficiently small neighbourhood of the point  $x$ , i.e. the whole graph is above the tangent hyperplane in a sufficiently small neighbourhood. In the case of negative definiteness, it is the other way round.

Finally, when the Hessian is indefinite, the graph of the function has values on both sides of the hyperplane. This happens, in general, along objects of lower dimensions in the tangent hyperplane, so there is no straightforward generalization of inflection points.

**8.1.19. The differential of mappings.** The concepts of derivative and differential can be easily extended to mappings  $F : E_n \rightarrow E_m$ . Having selected the Cartesian coordinate system on both sides, this mapping is an ordinary  $m$ -tuple

$$F(x_1, \dots, x_n) = (f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n))$$

of functions  $f_i : E_n \rightarrow \mathbb{R}$ .  $F$  is defined to be a *differentiable* or  *$k$ -times differentiable mapping* if and only if the corresponding property is shared by all the functions  $f_1, \dots, f_m$ .

The differentials  $df_i(x)$  of the particular functions  $f_i$  give a linear approximation of the increments of their values for the mapping  $F$ . Therefore, we can expect that they also give a coordinate expression of the linear mapping  $D^1F(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$  between the modelling spaces which linearly approximates the increments of the mapping  $F$ .

DIFFERENTIAL AND JACOBI MATRIX

Consider a differentiable mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  with components  $(f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n))$  and  $x$  in its domain. The matrix

$$D^1F(x) = \begin{pmatrix} df_1(x) \\ df_2(x) \\ \vdots \\ df_m(x) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} (x)$$

is called the *Jacobi matrix of the mapping  $F$*  at the point  $x$ .

The linear mapping  $D^1F(x)$  defined on the increments  $v = (v_1, \dots, v_n)$  by the Jacobi matrix is called the *differential of the mapping  $F$*  at a point  $x$  in the domain if and only if

$$\lim_{v \rightarrow 0} \frac{1}{\|v\|} \left( F(x + v) - F(x) - D^1F(x)(v) \right) = 0.$$

Recall that the definition of Euclidean distance guarantees that the limits of values in  $E_n$  exist if and only if the





the function on the given ellipsoid, while the other two are minima.  $\square$

**Remark.** Note that we have used the variable  $k$  instead of  $\lambda$  from the theorem 8.1.29.

**8.H.4.** Decide whether the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = z - xy^2$  has any minima and maxima on the sphere

$$x^2 + y^2 + z^2 = 1.$$

If so, determine them.

**Solution.** We are looking for solutions of the system

$$\begin{aligned} kx &= -y^2, \\ ky &= -2xy, \\ kz &= 1. \end{aligned}$$

The second equation implies that either  $y = 0$  or  $x = -\frac{k}{2}$ . The first possibility leads to the points  $[0, 0, 1], [0, 0, -1]$ . The second one cannot be satisfied. Note that because of the third equation  $k \neq 0$  and substituting into the equation of the sphere, we get the equation

$$\frac{k^2}{4} + \frac{k^2}{2} + \frac{1}{k^2} = 1,$$

which has no solution in real numbers (it is a quadratic equation in  $k^2$  with the negative discriminant). The function has a maximum and minimum, respectively, at the two computed points on the given sphere.  $\square$

**8.H.5.** Determine whether the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = xyz$ , has any extrema on the ellipsoid given by the equation

$$g(x, y, z) = kx^2 + ly^2 + z^2 = 1, \quad k, l \in \mathbb{R}^+.$$

If so, calculate them.

**Solution.** First, we build the equations which must be satisfied by the stationary points of the given function on the ellipsoid:

$$\begin{aligned} \frac{\partial g}{\partial x} &= \lambda \frac{\partial f}{\partial x} : yz = 2\lambda kx, \\ \frac{\partial g}{\partial y} &= \lambda \frac{\partial f}{\partial y} : xz = 2\lambda ly, \\ \frac{\partial g}{\partial z} &= \lambda \frac{\partial f}{\partial z} : xy = 2\lambda z. \end{aligned}$$

We can easily see that the equation can only be satisfied by a triple of non-zero numbers. Dividing pairs of equations and substituting into the ellipse's equation, we get eight solutions, namely the stationary points  $x = \pm \frac{1}{\sqrt{3k}}, y = \pm \frac{1}{\sqrt{3l}}, z = \pm \frac{1}{\sqrt{3}}$ . However, the function  $f$  takes on only two distinct values at these eight points. Since it is continuous and the given ellipsoid is compact,  $f$  must have both a maximum and minimum on it. Moreover, since both  $f$  and  $g$  are continuously differentiable, these extrema must occur at stationary points. Therefore, it must be that four of the computed stationary points are local maxima of the function (of value  $\frac{1}{3\sqrt{3kl}}$ ) and the other four are minima (of value  $-\frac{1}{3\sqrt{3kl}}$ ).  $\square$

limits of the particular coordinate components do. Direct application of Theorem 8.1.6 about the existence of the differential for functions of  $n$  variables to the particular coordinate functions of the mapping  $F$  thus leads to the following generalization (prove this in detail by yourselves!):

EXISTENCE OF THE DIFFERENTIAL

**Corollary.** Let  $F : E_n \rightarrow E_m$  be a mapping such that all of its coordinate functions have continuous partial derivatives in a neighbourhood of a point  $x \in E_n$ . Then the differential  $D^1F(x)$  exists, and it is given by the Jacobi matrix  $D^1F(x)$ .

**8.1.20. Lipschitz continuity.** Continuous differentiability of mappings allows good control on their variability in the following sense. Assume the estimates of the difference  $F(y) - F(x)$  for all  $x$  and  $y$  from a convex compact subset  $K$  in the domain of  $F$  are of interest. Applying the Taylor's theorem with remainder in order one on each of the components of  $F = (f_1, \dots, f_n)$  separately gives the estimate (write  $v = y - x$ )

$$\begin{aligned} \|F(y) - F(x)\|^2 &= \sum_{i=1}^m |f_i(y) - f_i(x)|^2 \\ &= \sum_{i=1}^m |D^1f_i(x + \theta_i v)(v)|^2 \\ &= \sum_{i=1}^m \sum_{j=1}^n \left| \frac{\partial f_i}{\partial x_j}(x + \theta_i v) v_j \right|^2 \\ &\leq \left( \max_{z \in K, i, j} \left| \frac{\partial f_i}{\partial x_j}(z) \right|^2 \right) nm \|v\|^2 = C^2 \|v\|^2 \end{aligned}$$

for an appropriate constant  $C \geq 0$ . The fact that continuous functions are bounded over each compact set is used.

This is the property of *Lipschitz continuity* of  $F$  on the compact set  $K$ :

$$\|F(y) - F(x)\| \leq C \|y - x\|, \quad \text{for all } x, y \in K$$

which was considered in 7.3.14 in the end of chapter 7.

**Proposition.** Each continuously differentiable mapping  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is Lipschitz continuous over convex compact sets.

**8.1.21. Differential of composite mappings.** The following theorem formulates a very useful generalization of the *chain rule* for univariate functions. Except for the concept of the differential itself, which is mildly complicated, it is actually the same as the one already seen in the case of one variable.

The Jacobi matrix for univariate functions is a single number, namely the derivative of the function at a given point, so the multiplication of Jacobi matrices is simply the multiplication of the derivatives of the outer and inner components of the function. There is, of course, another special case: the formula derived and used several times for the derivative of a composition of multivariate functions with curves. There, the differential is the linear form expressed via the partial derivatives of the outer components, evaluated on the vector of the

**8.H.6.** Determine the global extrema of the function

$$f(x, y) = x^2 - 2y^2 + 4xy - 6x - 1$$

on the set of points  $[x, y]$  that satisfy the inequalities

$$(1) \quad x \geq 0, \quad y \geq 0, \quad y \leq -x + 3.$$

**Solution.** We are given a polynomial with continuous partial derivatives on a compact (i. e. closed and bounded) set. Such a function necessarily has both a minimum and a maximum on this set, and this can happen only at stationary points or on the boundary. Therefore, it suffices to find stationary points inside the set and the ones on a finite number of open (or singleton) parts of the boundary, then evaluate  $f$  at these points and choose the least and the greatest values. Notice that the set of points determined by the inequalities (1) is clearly a triangle with vertices at  $[0, 0]$ ,  $[3, 0]$ ,  $[0, 3]$ .

Let us determine the stationary points inside this triangle as the solution of the equations  $f_x = 0$ ,  $f_y = 0$ . Since

$$f_x(x, y) = 2x + 4y - 6, \quad f_y(x, y) = 4x - 4y,$$

these equations are satisfied only by the point  $[1, 1]$ . The boundary suggests itself to be expressed as the union of three line segments given by the choice of pairs of vertices. First, we consider  $x = 0$ ,  $y \in [0, 3]$ , when  $f(x, y) = -2y^2 - 1$ . However, we know the graph of this (univariate) function on the interval  $[0, 3]$  It is thus not difficult to find the points at which global extrema occur. They are the marginal points  $[0, 0]$ ,  $[0, 3]$ . Similarly, we can consider  $y = 0$ ,  $x \in [0, 3]$ , also obtaining the marginal points  $[0, 0]$ ,  $[3, 0]$ . Finally, we get to the line segment  $y = -x + 3$ ,  $x \in [0, 3]$ . Making some rearrangements, we get

$$f(x, y) = f(x, -x + 3) = -5x^2 + 18x - 19, \quad x \in [0, 3].$$

We thus need to find the stationary points of the polynomial  $p(x) = -5x^2 + 18x - 19$  from the interval  $[0, 3]$ . The equation  $p'(x) = 0$ , i. e.,  $-10x + 18 = 0$ , is satisfied by  $x = 9/5$ . This means that in the last case, we obtained one more point (besides the marginal points), namely  $[9/5, 6/5]$ , where a global extremum may occur. Altogether, we have these points as “suspects”:

$$[1, 1], \quad [0, 0], \quad [0, 3], \quad [3, 0], \quad \left[\frac{9}{5}, \frac{6}{5}\right]$$

with function values

$$-4, \quad -1, \quad -19, \quad -10, \quad -\frac{14}{5},$$

respectively. We can see that the function  $f$  takes on the greatest value  $-1$  at the point  $[0, 0]$  and the least value  $-19$  at the point  $[0, 3]$ .  $\square$

**8.H.7.** Determine whether the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = y^2z$  has any extrema on the line segment given by the equations  $2x + y + z = 1$ ,  $x - y + 2z = 0$  and the constraint  $x \in [-1, 2]$ . If so, find these extrema and determine their types. Justify all of your decisions.

**Solution.** We are looking for the extrema of a continuous function on a compact set. Therefore, the function must have both a minimum and a maximum on this set, and this will happen either at the marginal points of the segment or at those

derivative of the inner component, again given by the product of the one line (the form) and one column (the vector).

THE CHAIN RULE

**Theorem.** Let  $F : E_n \rightarrow E_m$  and  $G : E_m \rightarrow E_r$  be two differentiable mappings, where the domain of  $G$  contains the whole image of  $F$ . Then, the composite mapping  $G \circ F$  is also differentiable, and its differential at any point  $x$  in the domain of  $F$  is given by the composition of differentials

$$D^1(G \circ F)(x) = D^1G(F(x)) \circ D^1F(x).$$

The Jacobi matrix on the left hand side is the product of the corresponding Jacobi matrices on the right hand side.

**PROOF.** In paragraph 8.1.6 and in the proof of Taylor’s theorem, it was derived how the differentiation of mappings composed of functions and curves behaves. This proved the theorem in the special case of  $n = r = 1$ . The general case can be proved analogously, one just has to work with more vectors.

Fix an arbitrary increment  $v$  and calculate the directional derivative for the composition  $G \circ F$  at a point  $x \in E_n$ . This means to determine the differentials for the particular coordinate functions of the mapping  $G$  composed with  $F$ . To simplify, write  $g \circ F$  for any one of them.

$$d_v(g \circ F)(x) = \lim_{t \rightarrow 0} \frac{1}{t} (g(F(x + tv)) - g(F(x))).$$

The expression in parentheses can, from the definition of the differential of  $g$ , be expressed as

$$g(F(x + tv)) - g(F(x)) = dg(F(x))(F(x + tv) - F(x)) + \alpha(F(x + tv) - F(x)),$$

where  $\alpha$  is a function defined on a neighbourhood of the point  $F(x)$  which is continuous and  $\lim_{v \rightarrow 0} \frac{1}{\|v\|} \alpha(v) = 0$ . Substitution into the equality for the directional derivative yields

$$\begin{aligned} d_v(g \circ F)(x) &= \lim_{t \rightarrow 0} \frac{1}{t} \left( dg(F(x))(F(x + tv) - F(x)) \right. \\ &\quad \left. + \alpha(F(x + tv) - F(x)) \right) \\ &= dg(F(x)) \left( \lim_{t \rightarrow 0} \frac{1}{t} (F(x + tv) - F(x)) \right) \\ &\quad + \lim_{t \rightarrow 0} \frac{1}{t} \left( \alpha(F(x + tv) - F(x)) \right) \\ &= dg(F(x)) \circ D^1F(x)(v) + 0. \end{aligned}$$

The fact that linear mappings between finite-dimensional spaces are always continuous was used. In the last step the Lipschitz continuity of  $F$ , i.e.  $\|F(x+tv) - F(x)\| \leq C\|v\||t|$  was exploited, and the properties of the function  $\alpha$ .

So the theorem for the particular functions  $g_1, \dots, g_r$  of the mapping  $G$  is proved. The theorem in general now follows from the definition of matrix multiplication and its links to linear mappings. Think about all details!  $\square$

where the gradient of the examined function is a linear combination of the gradients of the functions that give the constraints. First, let us look for the points which satisfy the gradient condition:

$$\begin{aligned} 0 &= 2k + l, \\ 2yz &= k - l, \\ y^2 &= k + 2l, \\ 2x + y + z &= 1, \\ x - y + 2z &= 0. \end{aligned}$$

The solution of the system is  $[x, y, z] = [\frac{2}{3}, 0, -\frac{1}{3}]$  and  $[x, y, z] = [\frac{4}{9}, \frac{2}{9}, -\frac{1}{9}]$  (of course, the variables  $k$  and  $l$  can also be computed, but we are not interested in them). The marginal points of the given line segment are  $[-1, \frac{5}{3}, \frac{4}{3}]$  and  $[2, -\frac{4}{3}, -\frac{5}{3}]$ . Considering these four points, the function takes on the greatest value at the first marginal point ( $f(x, y, z) = \frac{100}{27}$ ), which is its maximum on the given segment, and it takes the least value at the second marginal point ( $f(x, y, z) = -\frac{80}{27}$ ), which is thus its minimum there.  $\square$

**8.H.8.** Find the maximal and minimal values of the polynomial

$$p(x, y) = 4x^3 - 3x - 4y^3 + 9y$$

on the set

$$M = \{[x, y] \in \mathbb{R}^2; x^2 + y^2 \leq 1\}.$$

**Solution.** This is again the case of a polynomial on a compact set; therefore, we can restrict our attention to stationary points inside or on the boundary of  $M$  and the “marginal” points on the boundary of  $M$ . However, the only solutions of the equations

$$p_x(x, y) = 12x^2 - 3 = 0, \quad p_y(x, y) = -12y^2 + 9 = 0$$

are the points

$$\left[\frac{1}{2}, \frac{\sqrt{3}}{2}\right], \quad \left[\frac{1}{2}, -\frac{\sqrt{3}}{2}\right], \quad \left[-\frac{1}{2}, \frac{\sqrt{3}}{2}\right], \quad \left[-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right],$$

which are all on the boundary of  $M$ . This means that  $p$  has no extremum inside  $M$ . Now, it suffices to find the maximum and minimum of  $p$  on the unit circle  $k : x^2 + y^2 = 1$ . The circle  $k$  can be expressed parametrically as

$$x = \cos t, \quad y = \sin t, \quad t \in [-\pi, \pi].$$

Thus, instead of looking for the extrema of  $p$  on  $M$ , we are now seeking the extrema of the function

$$f(t) := p(\cos t, \sin t) = 4 \cos^3 t - 3 \cos t - 4 \sin^3 t + 9 \sin t$$

on the interval  $[-\pi, \pi]$ . For  $t \in [-\pi, \pi]$ , we have

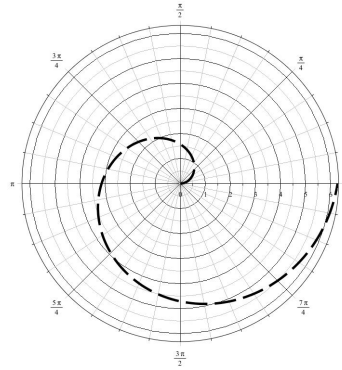
$$f'(t) = -12 \cos^2 t \sin t + 3 \sin t - 12 \sin^2 t \cos t + 9 \cos t,$$

In order to determine the stationary points, we must express the function  $f'$  in a form from which we will be able to calculate the intersection of its graph with the  $x$ -axis. To this purpose, we will use the identity

$$\frac{1}{\cos^2 t} = 1 + \tan^2 t,$$

**8.1.22. Transformation of coordinates.** A mapping  $F : E_n \rightarrow E_n$  which has an inverse mapping  $G : E_n \rightarrow E_n$  defined on the entire image of  $F$  is called a *transformation*. Such a mapping can be perceived as a change of coordinates. It is usually required that both  $F$  and  $G$  be (continuously) differentiable mappings.

Just as in the case of vector spaces, the choice of “point of view”, i.e. the choice of coordinates, can simplify or deteriorate comprehension of the examined object. The change of coordinates is now being discussed in a much more general form than in the case of affine mappings in the fourth chapter. Sometimes, the term “curvilinear coordinates” is used in this general sense. An illustrative example is the change of the most usual coordinates in the plane to polar coordinates. That is, the position of a point  $P$  is given by its distance  $r = \sqrt{x^2 + y^2}$  from the origin and the angle  $\varphi = \arctan(y/x)$  between the ray from the origin to it and the  $x$ -axis (if  $x \neq 0$ ). Notice, this is just the transformation between the algebraic and geometric form of a complex number.



The illustration shows the the “line”  $r = \varphi$  drawn in the Cartesian coordinates.

The change from the polar coordinates to the standard ones is

$$P_{\text{polar}} = (r, \varphi) \mapsto (r \cos \varphi, r \sin \varphi) = P_{\text{Cartesian}}$$

It is apparent that it is necessary to limit the polar coordinates to an appropriate subset of points  $(r, \varphi)$  in the plane so that the inverse mapping would exist. The Cartesian image of lines in polar coordinates with constant coordinates  $r$  or  $\varphi$  is also shown in the illustration above.

Let us discuss an example how to deal with the concept of transformation and the theorem about differentiation of composite mappings. The inverse to the above is the transformation  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  (for instance, on the domain of all points in the first quadrant except for the points having  $x = 0$ ):



$$r = \sqrt{x^2 + y^2}, \quad \varphi = \arctan \frac{y}{x}.$$

Consider now the function  $g_t : E_2 \rightarrow \mathbb{R}$ , with free parameter  $t \in \mathbb{R}$ ,

$$g(r, \varphi, t) = \sin(r - t)$$

which is valid provided both sides are well-defined. We get

$$f'(t) = \cos^3 t [-12 \operatorname{tg} t + 3 (\operatorname{tg} t + \operatorname{tg}^3 t) - 12 \operatorname{tg}^2 t + 9 (1 + \operatorname{tg}^2 t)]$$

for  $t \in [-\pi, \pi]$  with  $\cos t \neq 0$ . However, this condition does not exclude any stationary points since  $\sin t \neq 0$  if  $\cos t = 0$ . Therefore, the stationary points of  $f$  are those points  $t \in [-\pi, \pi]$  for which

$$-4 \operatorname{tg} t + \operatorname{tg} t + \operatorname{tg}^3 t - 4 \operatorname{tg}^2 t + 3 + 3 \operatorname{tg}^2 t = 0.$$

The substitution  $s = \operatorname{tg} t$  leads to

$$s^3 - s^2 - 3s + 3 = 0, \quad \text{i. e.} \quad (s - 1)(s - \sqrt{3})(s + \sqrt{3}) = 0.$$

Then, the values

$$s = 1, \quad s = \sqrt{3}, \quad s = -\sqrt{3}$$

respectively correspond to

$$t \in \{-\frac{3}{4}\pi, \frac{1}{4}\pi\}, \quad t \in \{-\frac{2}{3}\pi, \frac{1}{3}\pi\}, \quad t \in \{-\frac{1}{3}\pi, \frac{2}{3}\pi\}.$$

Now, we evaluate the function  $f$  at each of these points as well as at the marginal points  $t = -\pi, t = \pi$ . Sorting them, we get

$$\begin{aligned} f(-\frac{1}{3}\pi) &= -1 - 3\sqrt{3} < f(-\frac{3}{4}\pi) = -3\sqrt{2} < \\ & f(-\frac{2}{3}\pi) = 1 - 3\sqrt{3} < -1, \\ & f(-\pi) = f(\pi) = -1 < 0, \\ f(\frac{2}{3}\pi) &= 1 + 3\sqrt{3} > f(\frac{1}{4}\pi) = 3\sqrt{2} > f(\frac{1}{3}\pi) = \\ & -1 + 3\sqrt{3} > 0. \end{aligned}$$

Therefore, the global minimum of the function  $f$  is at the point  $t = -\pi/3$ , while the global maximum is at  $t = 2\pi/3$ .

Now, let us get back to the original function  $p$ . Since we know the values  $\cos(-\frac{1}{3}\pi) = \frac{1}{2}, \sin(-\frac{1}{3}\pi) = -\frac{\sqrt{3}}{2}, \cos(\frac{2}{3}\pi) = -\frac{1}{2}, \sin(\frac{2}{3}\pi) = \frac{\sqrt{3}}{2}$ , we can deduce that the polynomial  $p$  takes on the minimal value  $-1 - 3\sqrt{3}$  (the same as  $f$ , of course) at the point  $[1/2, -\sqrt{3}/2]$  and the maximal value  $1 + 3\sqrt{3}$  at  $[-1/2, \sqrt{3}/2]$ .  $\square$

**8.H.9.** At which points does the function

$$f(x, y) = x^2 - 4x + y^2$$

take on global extrema on the set  $M : |x| + |y| \leq 1$ ?

**Solution.** Expressing  $f$  in the form

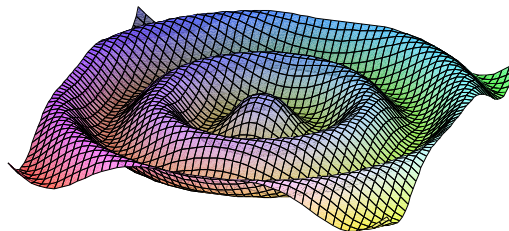
$$f(x, y) = (x - 2)^2 - 4 + y^2,$$

we can see that the global maximum and minimum occur at the same points as for the function

$$g(x, y) := \sqrt{(x - 2)^2 + y^2}, \quad [x, y] \in M,$$

since neither shifting the function nor applying the increasing function  $v = \sqrt{u}$  for  $u \geq 0$  changes the points of extrema (of course, they can change their values). However, we know that the function  $g$  gives the distance of a point  $[x, y]$  from the point  $[2, 0]$ . Since the set  $M$  is clearly a square with vertices  $[1, 0], [0, 1], [-1, 0], [0, -1]$ , the point of  $M$  that is closest to  $[2, 0]$  is the vertex  $[1, 0]$ , while the most distant one is  $[-1, 0]$ . Altogether, we have obtained that the minimal value of  $f$  occurs at the point  $[1, 0]$  and the maximal one at  $[-1, 0]$ .  $\square$

in polar coordinates. Such a function can approximate the waves on a water surface after a point impulse in the origin at the time  $t$ , see the illustration (there,  $t = -\pi/2$ ). While it was easy to define the function in polar coordinates, it would have been much harder to guess with Cartesian coordinates.



Compute the derivative of this function in Cartesian coordinates. Using the theorem,

$$\begin{aligned} \frac{\partial g}{\partial x}(x, y, t) &= \frac{\partial g}{\partial r}(r, \varphi) \frac{\partial r}{\partial x}(x, y) + \frac{\partial g}{\partial \varphi}(r, \varphi) \frac{\partial \varphi}{\partial x}(x, y) \\ &= \cos(\sqrt{x^2 + y^2} - t) \frac{x}{\sqrt{x^2 + y^2}} + 0 \end{aligned}$$

and, similarly,

$$\begin{aligned} \frac{\partial g}{\partial y}(x, y, t) &= \frac{\partial g}{\partial r}(r, \varphi) \frac{\partial r}{\partial y}(x, y) + \frac{\partial g}{\partial \varphi}(r, \varphi) \frac{\partial \varphi}{\partial y}(x, y) \\ &= \cos(\sqrt{x^2 + y^2} - t) \frac{y}{\sqrt{x^2 + y^2}}. \end{aligned}$$

**8.1.23. The inverse mapping theorem.** If the first derivative of a differentiable univariate function is non-zero, its sign determines whether the function is increasing or decreasing. Then, the function has this property in a neighbourhood of the point in question, and so an inverse function exists in the selected neighbourhood. The derivative of the inverse function  $f^{-1}$  is then the reciprocal value of the derivative of the function  $f$  (i.e. the inverse with respect to multiplication of real numbers). For higher dimensions there is the analogous result:



#### THE INVERSE MAPPING THEOREM

**Theorem.** Let  $F : E_n \rightarrow E_n$  be a differentiable mapping on a neighbourhood of a point  $x_0 \in E_n$ , and let the Jacobi matrix  $D^1 F(x_0)$  be invertible.

Then in some neighbourhood of  $y_0 = F(x_0)$ , the inverse mapping  $F^{-1}$  exists, it is differentiable, and its differential at the point  $F(x_0)$  is the inverse mapping to the differential  $D^1 F(x_0)$ .

Hence,  $D^1(F^{-1})(F(x_0))$  is given by the inverse matrix to the Jacobi matrix of the mapping  $F$  at the point  $x_0$ .

Interpreting this situation for a mapping  $E_1 \rightarrow E_1$  and linear mappings  $\mathbb{R} \rightarrow \mathbb{R}$  as their differentials, the nonvanishing is a necessary and sufficient condition for the differential

**8.H.10.** Compute the local extrema of the function  $y = f(x)$  given implicitly by the equation

$$3x^2 + 2xy + x = y^2 + 3y + \frac{5}{4}, \quad [x, y] \in \mathbb{R}^2 \setminus \left\{ \left[ x, x - \frac{3}{2} \right]; x \in \mathbb{R} \right\}.$$

**Solution.** In accordance with the theoretical part (see 8.1.25), let us denote

$$F(x, y) = 3x^2 + 2xy + x - y^2 - 3y - \frac{5}{4}, \quad [x, y] \in \mathbb{R}^2 \setminus \left\{ \left[ x, x - \frac{3}{2} \right]; x \in \mathbb{R} \right\}$$

and calculate the derivative

$$y' = f'(x) = -\frac{F_x(x, y)}{F_y(x, y)} = -\frac{6x+2y+1}{2x-2y-3}.$$

We can see that this derivative is continuous on the whole set in question. In particular, the function  $f$  is defined implicitly on this set (the denominator is non-zero).

A local extremum may occur only for those  $x, y$  which satisfy  $y' = 0$ , i. e.,  $6x + 2y + 1 = 0$ . Substituting  $y = -3x - 1/2$  into the equation  $F(x, y) = 0$ , we obtain  $-12x^2 + 6x = 0$ , which leads to

$$[x, y] = \left[ 0, -\frac{1}{2} \right], \quad [x, y] = \left[ \frac{1}{2}, -2 \right].$$

We can also easily compute that

$$y'' = (y')' = -\frac{(6+2y')(2x-2y-3) - (6x+2y+1)(2-2y')}{(2x-2y-3)^2}.$$

Substituting  $x = 0, y = -1/2, y' = 0$  and  $x = 1/2, y = -2, y' = 0$ , we obtain

$$y'' = -\frac{6(-2)-0}{4} > 0 \quad \text{for } [x, y] = \left[ 0, -\frac{1}{2} \right]$$

and

$$y'' = -\frac{6(+2)-0}{4} < 0 \quad \text{for } [x, y] = \left[ \frac{1}{2}, -2 \right].$$

We have thus proved that the implicitly given function has a strict local minimum at the point  $x = 0$  and a strict local maximum at  $x = 1/2$ .  $\square$

**8.H.11.** Find the local extrema of the function  $z = f(x, y)$  given on the maximum possible set by the equation

$$(1) \quad x^2 + y^2 + z^2 - xz - yz + 2x + 2y + 2z - 2 = 0.$$

**Solution.** Differentiating (1) with respect to  $x$  and  $y$  gives

$$2x + 2zz_x - z - xz_x - yz_x + 2 + 2z_x = 0,$$

$$2y + 2zz_y - xz_y - z - yz_y + 2 + 2z_y = 0.$$

Hence we get that

$$(2) \quad \begin{aligned} z_x = f_x(x, y) &= \frac{z - 2x - 2}{2z - x - y + 2}, \\ z_y = f_y(x, y) &= \frac{z - 2y - 2}{2z - x - y + 2}. \end{aligned}$$

We can notice that the partial derivatives are continuous at all points where the function  $f$  is defined. This implies that the local extrema can occur only at stationary points. These points satisfy

$$z_x = 0, \quad \text{i. e.} \quad z - 2x - 2 = 0,$$

$$z_y = 0, \quad \text{i. e.} \quad z - 2y - 2 = 0.$$

We have thus two equations, which allow us to express the dependency of  $x$  and  $y$  on  $z$ . Substituting into (1), we obtain the points

$$[x, y, z] = \left[ -3 + \sqrt{6}, -3 + \sqrt{6}, -4 + 2\sqrt{6} \right],$$

to be invertible as a linear mapping. In general finite dimensions, the non-generacy of the differential is the adequate concept.

**PROOF.** First, verify that the theorem makes sense and is as expected. If it is supposed that the inverse mapping exists and is differentiable at  $F(x_0)$ , then differentiating the composite mapping  $F^{-1} \circ F$  enforces the formula



$$\text{id}_{\mathbb{R}^n} = D^1(F^{-1} \circ F)(x_0) = D^1(F^{-1}) \circ D^1F(x_0),$$

which verifies the formula at the conclusion of the theorem. Therefore, it is known at the beginning which differential for  $F^{-1}$  to find.

Next, suppose that the inverse mapping  $F^{-1}$  exists in a neighbourhood of the point  $F(x_0)$  and that it is continuous. Since  $F$  is differentiable in a neighbourhood of  $x_0$ , it follows that



$$(1) \quad F(x) - F(x_0) - D^1F(x_0)(x - x_0) = \alpha(x - x_0)$$

with function  $\alpha : \mathbb{R}^n \rightarrow 0$  satisfying  $\lim_{v \rightarrow 0} \frac{1}{\|v\|} \alpha(v) = 0$ . To verify the approximation properties of the linear mapping  $(D^1F(x_0))^{-1}$ , it suffices to calculate the following limit for  $y = F(x)$  approaching  $y_0 = F(x_0)$ :

$$\lim_{y \rightarrow y_0} \frac{1}{\|y - y_0\|} (F^{-1}(y) - F^{-1}(y_0) - (D^1F(x_0))^{-1}(y - y_0)).$$

Substituting (1) for  $y - y_0$  into the latter equality yields

$$\begin{aligned} & \lim_{y \rightarrow y_0} \frac{1}{\|y - y_0\|} \left( x - x_0 \right. \\ & \quad \left. - (D^1F(x_0))^{-1}(D^1F(x_0)(x - x_0) + \alpha(x - x_0)) \right) \\ &= \lim_{y \rightarrow y_0} \frac{-1}{\|y - y_0\|} (D^1F(x_0))^{-1}(\alpha(x - x_0)) \\ &= (D^1F(x_0))^{-1} \lim_{y \rightarrow y_0} \frac{(-1)}{\|y - y_0\|} (\alpha(x - x_0)), \end{aligned}$$

where the last equality follows from the fact that linear mappings between finite-dimensional spaces are always continuous. Hence performing this linear mapping commutes with the limit process.

The proof is almost finished. The limit at the end of the expression is, using the properties of  $\alpha$ , zero if the values  $\|F(x) - F(x_0)\|$  are greater than  $C\|x - x_0\|$  for some constant  $C > 0$ . This can be translated in terms of the inverse as  $C\|F^{-1}(y) - F^{-1}(y_0)\| \leq \|y - y_0\|$ , i.e.

$$\|F^{-1}(y) - F^{-1}(y_0)\| \leq D\|y - y_0\|$$

for the constant  $D = C^{-1} > 0$ . This is Lipschitz continuity, which is a stronger property than  $F^{-1}$  being continuous. So, now it remains “merely” to prove the existence of a Lipschitz-continuous inverse mapping to the mapping  $F$ .

$$[x, y, z] = [-3 - \sqrt{6}, -3 - \sqrt{6}, -4 - 2\sqrt{6}].$$

Now, we need the second derivatives in order to decide whether the local extrema really occur at the corresponding points. Differentiating  $z_x$  in (2), we obtain

$$z_{xx} = f_{xx}(x, y) = \frac{(z_x - 2)(2z - x - y + 2) - (z - 2x - 2)(2z_x - 1)}{(2z - x - y + 2)^2},$$

with respect to  $x$ , and

$$z_{xy} = f_{xy}(x, y) = \frac{z_y(2z - x - y + 2) - (z - 2x - 2)(2z_y - 1)}{(2z - x - y + 2)^2},$$

with respect to  $y$ . We need not calculate  $z_{yy}$  since the variables  $x$  and  $y$  are interchangeable in (1) (if we swap  $x$  and  $y$ , the equation is left unchanged). Moreover, the  $x$ - and  $y$ -coordinates of the considered points are the same; hence  $z_{xx} = z_{yy}$ . Now, we evaluate that at the stationary points:

$$f_{xx}(-3 + \sqrt{6}, -3 + \sqrt{6}) = f_{yy}(-3 + \sqrt{6}, -3 + \sqrt{6}) = -\frac{1}{\sqrt{6}},$$

$$f_{xy}(-3 + \sqrt{6}, -3 + \sqrt{6}) = f_{yx}(-3 + \sqrt{6}, -3 + \sqrt{6}) = 0,$$

$$f_{xx}(-3 - \sqrt{6}, -3 - \sqrt{6}) = f_{yy}(-3 - \sqrt{6}, -3 - \sqrt{6}) = \frac{1}{\sqrt{6}},$$

$$f_{xy}(-3 - \sqrt{6}, -3 - \sqrt{6}) = f_{yx}(-3 - \sqrt{6}, -3 - \sqrt{6}) = 0.$$

As for the Hessian, we have

$$Hf(-3 + \sqrt{6}, -3 + \sqrt{6}) = \begin{pmatrix} -\frac{1}{\sqrt{6}} & 0 \\ 0 & -\frac{1}{\sqrt{6}} \end{pmatrix},$$

$$Hf(-3 - \sqrt{6}, -3 - \sqrt{6}) = \begin{pmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

Apparently, the first Hessian is negative definite, while the second one is positive definite. This means that there is a strict local maximum of the function  $f$  at the point  $[-3 + \sqrt{6}, -3 + \sqrt{6}]$ , and there is a strict local minimum at the point  $[-3 - \sqrt{6}, -3 - \sqrt{6}]$ .  $\square$

**8.H.12.** Determine the strict local extrema of the function

$$f(x, y) = \frac{1}{x} + \frac{1}{y}, \quad x \neq 0, \quad y \neq 0$$

on the set of points that satisfy the equation  $\frac{1}{x^2} + \frac{1}{y^2} = 4$ .

**Solution.** Since both the function  $f$  and the function given implicitly by the equation  $\frac{1}{x^2} + \frac{1}{y^2} - 4 = 0$  have continuous partial derivatives of all orders on the set  $\mathbb{R}^2 \setminus \{[0, 0]\}$ , we should look for stationary points, i. e., for the solution of the equations  $L_x = 0, L_y = 0$  for

$$L(x, y, \lambda) = \frac{1}{x} + \frac{1}{y} - \lambda \left( \frac{1}{x^2} + \frac{1}{y^2} - 4 \right), \quad x \neq 0, \quad y \neq 0.$$

We thus get the equations

$$-\frac{1}{x^2} + \frac{2\lambda}{x^3} = 0, \quad -\frac{1}{y^2} + \frac{2\lambda}{y^3} = 0,$$

which lead to  $x = 2\lambda, y = 2\lambda$ . Considering the set of points in question, the constraint  $x = y$  gives the stationary points

$$(1) \quad \left[ \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2} \right], \quad \left[ -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} \right].$$

Now, let us examine the second differential of the function  $L$ . We can easily compute that



To simplify, reduce the general case slightly. Especially, without loss of generality, apply shifts of the coordinates by constant vectors. In particular, it can be assumed that  $x_0 = 0 \in \mathbb{R}^n, y_0 = F(x_0) = 0 \in \mathbb{R}^n$ . So assume this property of the mapping  $F$ .

Further, composing the mapping  $F$  with any linear mapping  $G$  yields a differentiable mapping again, and it is known how the differential changes. The choice  $G(y) = (D^1F(0))^{-1}(y)$  gives  $D^1(G \circ F)(0) = \text{id}_{\mathbb{R}^n}$  and thus we may assume that

$$D^1F(0) = \text{id}_{\mathbb{R}^n}.$$

With these assumptions, consider the mapping  $K(x) = F(x) - x$ . This mapping is also differentiable, and its differential at 0 is zero.

It is already known that each continuously differentiable mapping is Lipschitz continuous over every  $\delta$ -neighbourhood  $U_\delta$  of the origin (in its domain),

$$\|K(x) - K(y)\| \leq C\|x - y\|,$$

where  $C$  is bounded by the maximum of all absolute values of the partial derivatives in the Jacobi matrix of the mapping  $K$  in the neighbourhood  $U_\delta$ , cf. 8.1.20.

Since the differential of the mapping  $K$  at the point  $x_0 = 0$  is zero, one can, by selecting a sufficiently small neighbourhood  $U$  of the origin, achieve the bound

$$\|K(x) - K(y)\| \leq \frac{1}{2}\|x - y\|.$$

It follows by the triangle inequality that

$$\begin{aligned} \|x - y\| &= \|(F(x) - K(x)) - (F(y) - K(y))\| \\ &\leq \|F(x) - F(y)\| + \|K(x) - K(y)\| \\ &\leq \|F(x) - F(y)\| + \frac{1}{2}\|x - y\| \end{aligned}$$

and hence

$$\frac{1}{2}\|x - y\| \leq \|F(x) - F(y)\|.$$

With this estimate, if  $x \neq y$  are both in the neighbourhood  $U = U_\delta$ , then also  $F(x) \neq F(y)$ . Therefore, the mapping is bijective onto its image  $V = F(U)$ . Write  $F^{-1}$  for its inverse defined on  $V$ . For this mapping, the latter estimate says

$$\|F^{-1}(x) - F^{-1}(y)\| \leq 2\|x - y\|,$$

so this mapping is not only continuous (as we assumed in our first step in the proof), but also Lipschitz-continuous, as requested in the end of the previous part of the proof.

It could seem that the proof is complete, but this is not so.

To finish, it is necessary to show that the mapping  $F$  restricted to a sufficiently small neighbourhood  $U_\delta$  is not only bijective onto its image, but also that it maps open neighbourhoods of zero onto open neighbourhoods of zero.<sup>3</sup>



<sup>3</sup>In the literature, there are examples of mappings which continuously and bijectively map a line segment onto a square. So this is not an obvious requirement.

$$L_{xx} = \frac{2}{x^3} - \frac{6\lambda}{x^4}, \quad L_{xy} = 0, \quad L_{yy} = \frac{2}{y^3} - \frac{6\lambda}{y^4}, \quad x \neq 0, y \neq 0,$$

whence it follows that

$$d^2L(x, y) = \left(\frac{2}{x^3} - \frac{6\lambda}{x^4}\right) dx^2 + \left(\frac{2}{y^3} - \frac{6\lambda}{y^4}\right) dy^2.$$

Differentiating the constraint  $\frac{1}{x^2} + \frac{1}{y^2} = 4$ , we get

$$-\frac{2}{x^3} dx - \frac{2}{y^3} dy = 0, \quad \text{i. e.} \quad dy^2 = \frac{y^6}{x^6} dx^2.$$

Therefore,

$$d^2L(x, y) = \left[\frac{2}{x^3} - \frac{6\lambda}{x^4} + \left(\frac{2}{y^3} - \frac{6\lambda}{y^4}\right) \frac{y^6}{x^6}\right] dx^2.$$

In fact, we are considering a one-dimensional quadratic form whose positive (negative) definiteness at a stationary point means that there is a minimum (maximum) at that point. Realizing that the stationary points had  $x = 2\lambda$ ,  $y = 2\lambda$ , mere substitution yields

$$d^2L\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = -4\sqrt{2} dx^2, \quad d^2L\left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) = 4\sqrt{2} dx^2,$$

which means that there is a strict local maximum of the function  $f$  at the point  $[\sqrt{2}/2, \sqrt{2}/2]$ , while at the point  $[-\sqrt{2}/2, -\sqrt{2}/2]$ , there is a strict local minimum. The corresponding values are:

$$(2) \quad f\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = 2\sqrt{2}, \quad f\left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) = -2\sqrt{2}.$$

Now, we will demonstrate a quicker way how to obtain the result. We know (or we can easily calculate) the second partial derivatives of the function  $L$ , i. e., the Hessian with respect to the variables  $x$  and  $y$ :

$$HL(x, y) = \begin{pmatrix} \frac{2}{x^3} - \frac{6\lambda}{x^4} & 0 \\ 0 & \frac{2}{y^3} - \frac{6\lambda}{y^4} \end{pmatrix}.$$

The evaluation

$$HL\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = \begin{pmatrix} -2\sqrt{2} & 0 \\ 0 & -2\sqrt{2} \end{pmatrix},$$

$$HL\left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) = \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \end{pmatrix}$$

then tells us that the quadratic form is negative definite for the former stationary point (there is a strict local maximum) and positive definite for the latter one (there is a strict local minimum).

We should be aware of a potential trap in this “quicker” method in the case we obtain an indefinite form (matrix). Then, we cannot conclude that there is not an extremum at that point since as we have not included the constraint (which we did when computing  $d^2L$ ), we are considering a more general situation. The graph of the function  $f$  on the given set is a curve which can be defined as a univariate function. This must correspond to a one-dimensional quadratic form.  $\square$

Decrease the latter neighbourhood  $U = U_\delta$  so that the above estimates are true for the boundary of  $U$  as well and at the same time the Jacobi matrix of the mapping is invertible on all of  $U$ . This can be done since the determinant is a continuous mapping. Let  $B$  denote the boundary of the set  $U$ , that is, the corresponding sphere. Since  $B$  is compact and  $F$  is continuous, the function

$$\rho(x) = \|F(x)\|$$

achieves both the maximum and the minimum on  $B$ . Denote  $a = \frac{1}{2} \min_{x \in B} \rho(x)$  and consider any  $y \in \mathcal{O}_a(0)$  fixed. Of course,  $a > 0$  because  $x = 0$  is the only point with  $F(x) = 0$  within  $U_\delta$ . It is necessary to show that there is at least one  $x \in U$  such that  $y = F(x)$ , which completes the proof of the inverse mapping theorem.

For this purpose, consider the function ( $y$  is a fixed point)

$$h(x) = \|F(x) - y\|^2.$$

Again, the image  $h(U) \cup h(B)$  must have a minimum. This minimum cannot occur for  $x \in B$ .

Notice that  $F(0) = 0$ , hence  $h(0) = \|y\| < a$ . At the same time, the distance of  $y$  from  $F(x)$  for  $x \in B$  is at least  $a$  for all  $y \in \mathcal{O}_a(0)$  (since  $a$  is selected to be half the minimum of the magnitude of  $F(x)$  on the boundary). Therefore, the minimum occurs inside  $U$ , and it is a stationary point  $z$  of the function  $h$ . Fixing such  $z$ , means that for all  $j = 1, \dots, n$ ,

$$\frac{\partial h}{\partial x^j}(z) = \sum_{i=1}^n 2(f_i(z) - y_i) \frac{\partial f_i}{\partial x^j}(z) = 0.$$

This is a system of linear equations with variables  $\xi_i = f_i(z) - y_i$  and coefficients given by twice the Jacobi matrix  $D^1F(z)$ . In particular, for  $z \in U$ , such a system has a unique solution, and this is zero since the Jacobi matrix is invertible.

In this way the desired point  $x = z \in U$  is found, satisfying, for all  $i = 1, \dots, n$ , the equality  $f_i(z) = y_i$ , i.e.,  $F(z) = y$ .  $\square$

**8.1.24. The implicit functions.** The next goal is to employ the inverse mapping theorem for clarifying the properties of implicitly defined functions. To start, consider a differentiable function  $F(x, y)$  defined in the plane  $E_2$ , and look for those points  $(x, y)$  where  $F(x, y) = 0$ .

An example of this can be the usual (implicit) definition of straight lines and circles:

$$F(x, y) = ax + by + c = 0, \quad a, b, c \in \mathbb{R}$$

$$F(x, y) = (x - s)^2 + (y - t)^2 - r^2 = 0, \quad r > 0.$$

While in the first case, the relation between the quantities  $x$  and  $y$  can be expressed as the function (for  $b \neq 0$ )

$$y = f(x) = -\frac{a}{b}x - \frac{c}{b}$$

for all  $x$ ; in the other case, for any point  $(x_0, y_0)$  satisfying the equation of the circle and such that  $y_0 \neq t$  (these are the



**8.H.13.** Find the global extrema of the function

$$f(x, y) = \frac{1}{x} + \frac{1}{y}, \quad x \neq 0, y \neq 0$$

on the set of points that satisfy the equation  $\frac{1}{x^2} + \frac{1}{y^2} = 4$ .

**Solution.** This exercise is to illustrate that looking for global extrema may be much easier than for local ones (cf. the above exercise) even in the case when the function values are considered on an unbounded set. First, we would determine the stationary points (1) and the values (2) the same way as above. Let us emphasize that we are looking for the function's extrema on a set that is *not* compact, so we will not do with evaluating the function at the stationary points. The reason is that the function  $f$  may not have an extremum on the considered set – its range might be an open interval. However, we will show that this is not the case here.

Let us thus consider  $|x| \geq 10$ . We can realize that the equation  $\frac{1}{x^2} + \frac{1}{y^2} = 4$  can be satisfied only by those values  $y$  for which  $|y| \geq 1/2$ . We have thus obtained the bounds

$$-2\sqrt{2} < -\frac{1}{10} - 2 \leq f(x, y) \leq \frac{1}{10} + 2 < 2\sqrt{2}, \quad \text{if } |x| \geq 10.$$

At the same time, we have (interchanging  $x$  and  $y$  leads to the same task)

$$-2\sqrt{2} < -\frac{1}{10} - 2 \leq f(x, y) \leq \frac{1}{10} + 2 < 2\sqrt{2}, \quad \text{if } |y| \geq 10.$$

Hence we can see that the function  $f$  must have global extrema on the considered set, and this must happen inside the square  $ABCD$  with vertices  $A = [-10, -10]$ ,  $B = [10, -10]$ ,  $C = [10, 10]$ ,  $D = [-10, 10]$ .

The intersection of the “hundred times reduced” square with vertices at  $\tilde{A} = [-1/10, -1/10]$ ,  $\tilde{B} = [1/10, -1/10]$ ,  $\tilde{C} = [1/10, 1/10]$ ,  $\tilde{D} = [-1/10, 1/10]$  and the given set is clearly the empty set. Therefore, the global extrema are at points inside the compact set bounded by these two squares. Since  $f$  is continuously differentiable on this set, the global extrema can occur only at stationary points. We thus must have

$$f_{\max} = f\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right) = 2\sqrt{2}, \quad f_{\min} = f\left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) = -2\sqrt{2}.$$

□

**8.H.14.** Determine the maximal and minimal values of the function  $f(x, y, z) = xyz$  on the set  $M$  given by the conditions

$$x^2 + y^2 + z^2 = 1, \quad x + y + z = 0.$$

**Solution.** It is not hard to realize that  $M$  is a circle. However, for our problem, it is sufficient to know that  $M$  is compact, i. e. bounded (the first condition of the equation of the unit sphere) and closed (the set of solutions of the given equations is closed since if the equations are satisfied by all terms of a converging sequence, then it is satisfied by its limit as well). The function  $f$  as well as the constraint functions  $F(x, y, z) = x^2 + y^2 + z^2 - 1$ ,  $G(x, y, z) = x + y + z$

marginal points of the circle in the direction of the coordinate  $x$ ). There is a neighbourhood of the point  $x_0$  in which either

$$y = f(x) = t + \sqrt{(x-s)^2 - r},$$

or

$$y = f(x) = t - \sqrt{(x-s)^2 - r},$$

according to whether  $(x_0, y_0)$  belongs to the upper or lower semicircle.

If a diagram of the situation is drawn, the reason is clear: describing both the semicircles simultaneously by a single function  $y = f(x)$  is not possible. The boundary points of the interval  $[s-r, s+r]$  are even more interesting. They also satisfy the equation of the circle with  $y = t$ , yet  $F_y(s \pm r, t) = 0$ , which describes the position of the tangent line to the circle at these points, parallel to the  $y$ -axis. There are no neighbourhoods of these points in which the circle could be described as a function  $y = f(x)$ .

Moreover, the derivatives of the function  $y = f(x) = t + \sqrt{(x-s)^2 - r^2}$  can be easily expressed in terms of partial derivatives of the function  $F$ :

$$f'(x) = \frac{1}{2} \frac{2(x-s)}{\sqrt{r^2 - (x-s)^2}} = \frac{x-s}{y-t} = -\frac{F_x}{F_y}.$$

If the roles of the variables  $x$  and  $y$  are interchanged and a relation  $x = f(y)$  such that  $F(f(y), y) = 0$  is sought, then neighbourhoods of the points  $(s \pm r, t)$  are obtained with no problem. Notice that the partial derivative  $F_x$  is non-zero at these points.

So it is observed (though for only two examples): for a function  $F(x, y)$  and a point  $(a, b) \in E_2$  such that  $F(a, b) = 0$ , there is the unique function  $y = f(x)$  satisfying  $F(x, f(x)) = 0$  on some neighbourhood of  $x$  if  $F_y(a, b) \neq 0$ . In this case,  $f'(a) = -F_x(a, b)/F_y(a, b)$  can even be computed. We prove that in fact, this proposition is always true.

The last statement about derivatives can be remembered (and is quite comprehensible if things are properly understood) from the expression for the differential of the (constant) function  $g(x) = F(x, y(x))$  and the differential  $dy = f'(x)dx$

$$0 = dg = F_x dx + F_y dy = (F_x + F_y f'(x)) dx.$$

One can work analogously with the implicit expressions  $F(x, y, z) = 0$ , to look for a function  $g(x, y)$  such that  $F(x, y, g(x, y)) = 0$ . As an example, consider the function  $f(x, y) = x^2 + y^2$ , whose graph is the rotational paraboloid centered at the point  $(0, 0)$ . This can be defined implicitly by the equation

$$0 = F(x, y, z) = z - x^2 - y^2.$$

Before formulating the result for the general situation, notice which dimensions could/should appear in the problem. If it is desired to find, for this function  $F$ , a curve  $c(x) = (c_1(x), c_2(x))$  in the plane such that

$$F(x, c(x)) = F(x, c_1(x), c_2(x)) = 0,$$



have continuous partial derivatives of all orders (since they are polynomials). The Jacobi constraint matrix is

$$\begin{pmatrix} F_x(x, y, z) & F_y(x, y, z) & F_z(x, y, z) \\ G_x(x, y, z) & G_y(x, y, z) & G_z(x, y, z) \end{pmatrix} = \begin{pmatrix} 2x & 2y & 2z \\ 1 & 1 & 1 \end{pmatrix}.$$

Its rank is reduced (less than 2) if and only if the vector  $(2x, 2y, 2z)$  is a multiple of the vector  $(1, 1, 1)$ , which gives  $x = y = z$ , and thus  $x = y = z = 0$  (by the second constraint). However, the set  $M$  does contain the origin. Therefore, we may look for stationary points using the method of Lagrange multipliers. For

$$L(x, y, z, \lambda_1, \lambda_2) = xyz - \lambda_1(x^2 + y^2 + z^2 - 1) - \lambda_2(x + y + z),$$

the equations  $L_x = 0, L_y = 0, L_z = 0$  give

$$yz - 2\lambda_1 x - \lambda_2 = 0,$$

$$xz - 2\lambda_1 y - \lambda_2 = 0,$$

$$xy - 2\lambda_1 z - \lambda_2 = 0,$$

respectively. Subtracting the first equation from the second one and from the third one leads to

$$xz - yz - 2\lambda_1 y + 2\lambda_1 x = 0,$$

$$xy - yz - 2\lambda_1 z + 2\lambda_1 x = 0,$$

i. e.,

$$(x - y)(z + 2\lambda_1) = 0,$$

$$(x - z)(y + 2\lambda_1) = 0.$$

The last equations are satisfied in these four cases:

$$x = y, x = z; \quad x = y, y = -2\lambda_1;$$

$$z = -2\lambda_1, x = z; \quad z = -2\lambda_1, y = -2\lambda_1,$$

thus (including the constraint  $G = 0$ )

$$x = y = z = 0; \quad x = y = -2\lambda_1, z = 4\lambda_1;$$

$$x = z = -2\lambda_1, y = 4\lambda_1; \quad x = 4\lambda_1, y = z = -2\lambda_1.$$

Except for the first case (which clearly cannot happen), including the constraint  $F = 0$  yields

$$(4\lambda_1)^2 + (-2\lambda_1)^2 + (-2\lambda_1)^2 = 1, \quad \text{i. e.} \quad \lambda_1 = \pm \frac{1}{2\sqrt{6}}.$$

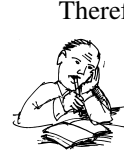
Altogether, we get the points

$$\left[-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}\right], \quad \left[-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right], \quad \left[\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right], \\ \left[\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right], \quad \left[\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right], \quad \left[-\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right].$$

We will not verify that these really are stationary points. The only important thing is that all stationary points are among these six.

We are looking for the global maximum and minimum of the continuous function  $f$  on the compact set  $M$ . However, the global extrema (we know they exist) can occur only at points of local extrema with respect to  $M$ . And the local extrema can occur only at the aforementioned points. Therefore, it suffices to evaluate the function  $f$  at these points. Thus

then this can be done (even for all initial conditions  $x = a$ ), yet the result is not unique for a given initial condition. It suffices to consider an arbitrary curve on the rotational paraboloid whose projection onto the first coordinate has a non-zero derivative. Then consider  $x$  to be the parameter of the curve, and  $c(x)$  to be its projection onto the plane  $yz$ .



Therefore, it is expected that one function of  $m + 1$  variables defines implicitly a hypersurface in  $\mathbb{R}^{m+1}$  which is to be expressed (at least locally) as the graph of a function of  $m$  variables. It can be anticipated that  $n$  functions of  $m + n$  variables define an intersection of  $n$  hypersurfaces in  $\mathbb{R}^{m+n}$ , which is expected as an “ $m$ -dimensional” object.

**8.1.25. The general theorem.** Consider a differentiable mapping

$$F = (f_1, \dots, f_n) : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^n.$$

The Jacobi matrix of this mapping has  $n$  rows and  $m + n$  columns. Write it symbolically as

$$D^1 F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} & \frac{\partial f_1}{\partial x_{m+1}} & \cdots & \frac{\partial f_1}{\partial x_{m+n}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} & \frac{\partial f_n}{\partial x_{m+1}} & \cdots & \frac{\partial f_n}{\partial x_{m+n}} \end{pmatrix} \\ = (D_x^1 F, D_y^1 F),$$

where  $(x_1, \dots, x_{m+n}) \in \mathbb{R}^{m+n}$  is written as  $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$ ,  $D_x^1 F$  is a matrix of  $n$  rows and the first  $m$  columns in the Jacobi matrix, while  $D_y^1 F$  is a square matrix of order  $n$ , with the remaining columns. The multidimensional analogy to the previous reasoning with the non-zero partial derivative with respect to  $y$  is the condition that the matrix  $D_y^1 F$  is invertible.

#### THE IMPLICIT MAPPING THEOREM

**Theorem.** Let  $F : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^n$  be a differentiable mapping in an open neighbourhood of a point  $(a, b) \in \mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$  at which  $F(a, b) = 0$ , and  $\det D_y^1 F \neq 0$ .

Then there exists a neighbourhood  $U$  of the point  $a \in \mathbb{R}^m$  and a unique differentiable mapping  $G : \mathbb{R}^m \rightarrow \mathbb{R}^n$  defined on  $U$ , with  $G(a) = b$  and such that  $F(x, G(x)) = 0$  for all  $x \in U$ .

Moreover, the Jacobi matrix  $D^1 G$  of the mapping  $G$  is, in the neighbourhood of the point  $a$ , given by the product of matrices

$$D^1 G(x) = -(D_y^1 F)^{-1}(x, G(x)) \cdot D_x^1 F(x, G(x)).$$



**PROOF.** For the sake of comprehensibility, first show the proof for the simplest case of the equation  $F(x, y) = 0$  with a function  $F$  of two variables. At first sight, it might look complicated, but this situation can be discussed in a way which can be extended for the general dimensions as in the theorem, almost without changes.

we find out that the wanted maximum is

$$\begin{aligned} f\left(-\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}\right) &= f\left(-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right) = \\ &= f\left(\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}\right) = \frac{1}{3\sqrt{6}}, \end{aligned}$$

while the minimum is

$$\begin{aligned} f\left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right) &= f\left(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) = \\ &= f\left(-\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) = -\frac{1}{3\sqrt{6}}. \end{aligned}$$

□

**8.H.15.** Find the extrema of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^2 + y^2 + z^2$ , on the plane  $x + y - z = 1$  and determine their types.

**Solution.** We can easily build the equations that describe the linear dependency between the normal to the constraint surface and the examined function:

$$x = k, \quad y = k, \quad z = -k, \quad k \in \mathbb{R}.$$

The only solution is the point  $[\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}]$ . Further, we can notice that the function is increasing in the direction of  $(1, -1, 0)$ , and this direction lies in the constraint plane. Therefore, the examined function has a minimum at this point.

**Another solution.** We will reduce this problem to finding the extrema of a two-variable function on  $\mathbb{R}^2$ . Since the constraint is linear, we can express  $z = x + y - 1$ . Substituting this into the given function then yields a real-valued function of two variables:  $f(x, y) = x^2 + y^2 + (x + y - 1)^2 = 2x^2 + 2xy + y^2 - 2x - 2y + 1$ . Setting both partial derivatives equal to zero, we get the linear equation

$$4x + 2y - 2 = 0, \quad 4y + 2x - 2 = 0,$$

whose only solution is the point  $[\frac{1}{3}, \frac{1}{3}]$ . Since it is a quadratic function with positive coefficients at the unknowns, it is unbounded on  $\mathbb{R}^2$ . Therefore, there is a (global) minimum at the obtained point. Then, we can get the corresponding point  $[\frac{1}{3}, \frac{1}{3}, -\frac{1}{3}]$  in the constraint plane from the linear dependency of  $z$ . □

**8.H.16.** Find the extrema of the function  $x + y : \mathbb{R}^3 \rightarrow \mathbb{R}$  on the circle given by the equations  $x + y + z = 1$  and  $x^2 + y^2 + z^2 = 4$ .

**Solution.** The “suspects” are those points which satisfy

$$(1, 1, 0) = k \cdot (1, 1, 1) + l \cdot (x, y, z), \quad k, l \in \mathbb{R}.$$

Clearly,  $x = y (= 1/l)$ . Substituting this into the equation of the circle then leads to the two solutions

$$\left[ \frac{1}{3} \pm \frac{\sqrt{22}}{6}, \frac{1}{3} \pm \frac{\sqrt{22}}{6}, \frac{1}{3} \mp \frac{\sqrt{22}}{3} \right].$$

Since every circle is compact, it suffices to examine the function values at these two points. We find out that there is a

Extend the function  $F$  to

$$\tilde{F} : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \quad (x, y) \mapsto (x, F(x, y)).$$

The Jacobi matrix of the mapping  $\tilde{F}$  is

$$D^1 \tilde{F}(x, y) = \begin{pmatrix} 1 & 0 \\ F_x(x, y) & F_y(x, y) \end{pmatrix}.$$

It follows from the assumption  $F_y(a, b) \neq 0$ , that the same also holds in a neighbourhood of the point  $(a, b)$ , so the function  $\tilde{F}$  is invertible in this neighbourhood, by the inverse mapping theorem. Therefore, there is a uniquely defined differentiable inverse mapping  $\tilde{F}^{-1}$  in a neighbourhood of the point  $(a, 0)$ .

Denote by  $\pi : \mathbb{R}^2 \rightarrow \mathbb{R}$  the projection onto the second coordinate, and consider the function  $f(x) = \pi \circ \tilde{F}^{-1}(x, 0)$ . This function is well-defined and differentiable. It must be verified that the expression

$$F(x, f(x)) = F(x, \pi(\tilde{F}^{-1}(x, 0)))$$

is zero in a neighbourhood of the point  $x = a$ . It follows directly from the definition of  $\tilde{F}(x, y) = (x, F(x, y))$  that its inverse is of the form  $\tilde{F}^{-1}(x, y) = (x, \pi \tilde{F}^{-1}(x, y))$ . Therefore, the previous calculation can be resumed:

$$\begin{aligned} F(x, f(x)) &= \pi(\tilde{F}(x, \pi(\tilde{F}^{-1}(x, 0)))) \\ &= \pi(\tilde{F}(\tilde{F}^{-1}(x, 0))) = \pi(x, 0) = 0. \end{aligned}$$

This proves the first part of the theorem, and it remains to compute the derivative of the function  $f(x)$ . This derivative can, once again, be obtained by invoking the inverse mapping theorem, using the matrix  $(D^1 \tilde{F})^{-1}$ .

The following equality is easily verified by multiplying the matrices. It can also be computed directly using the explicit formula for the inverse matrix in terms of the determinant and the algebraically adjoint matrix, see paragraph 2.2.11

$$\begin{pmatrix} 1 & 0 \\ F_x(x, y) & F_y(x, y) \end{pmatrix}^{-1} = (F_y(x, y))^{-1} \begin{pmatrix} F_y(x, y) & 0 \\ -F_x(x, y) & 1 \end{pmatrix}.$$

By the definition  $f(x) = \pi \tilde{F}^{-1}(x, 0)$ , and thus the first entry of the second row of this matrix is the derivative  $f'(x)$  with  $y = f(x)$ , i.e. the Jacobi matrix  $D^1 f$ . In this simple case, it is exactly the desired scalar  $-F_x(x, f(x))/F_y(x, f(x))$ .

The general proof is exactly the same, there is no need to change any of the formulae. We obtain the invertible mapping  $\tilde{F} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^{m+n}$  and define  $G(x) = \pi \tilde{F}^{-1}(x, 0)$ , where  $\pi : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^n$ ,  $\pi(x, y) = y$ . The same check as above reveals that  $F(x, G(x)) = 0$  as requested. Only in the last computation of the derivative of the function do the corresponding parts of the Jacobi matrix  $D_x^1 F$  and  $D_y^1 F$  appear, instead of the particular partial derivatives.

For the calculation of the Jacobi matrix of the mapping  $G$ , use the computation of the inverse matrix. This time the algebraic procedure from paragraph 2.2.11 is not very advantageous. It is better to be guided by the case in dimension



maximum of the considered function on the given circle at the former point and a minimum at the latter one.  $\square$

**8.H.17.** Find the extrema of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,  $f(x, y, z) = x^2 + y^2 + z^2$ , on the plane  $2x + y - z = 1$  and determine their types.  $\circ$

**8.H.18.** Find the maximum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = xy$  on the circle with radius 1 which is centered at the point  $[x_0, y_0] = [0, 1]$ .  $\circ$

**8.H.19.** Find the minimum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f = xy$  on the circle with radius 1 which is centered at the point  $[x_0, y_0] = [2, 0]$ .  $\circ$

**8.H.20.** Find the minimum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f = xy$  on the circle with radius 1 which is centered at the point  $[x_0, y_0] = [2, 0]$ .  $\circ$

**8.H.21.** Find the minimum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f = xy$  on the ellipse  $x^2 + 3y^2 = 1$ .  $\circ$

**8.H.22.** Find the minimum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f = x^2y$  on the circle with radius 1 which is centered at the point  $[x_0, y_0] = [0, 0]$ .  $\circ$

**8.H.23.** Find the maximum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x^3y$  on the circle  $x^2 + y^2 = 1$ .  $\circ$

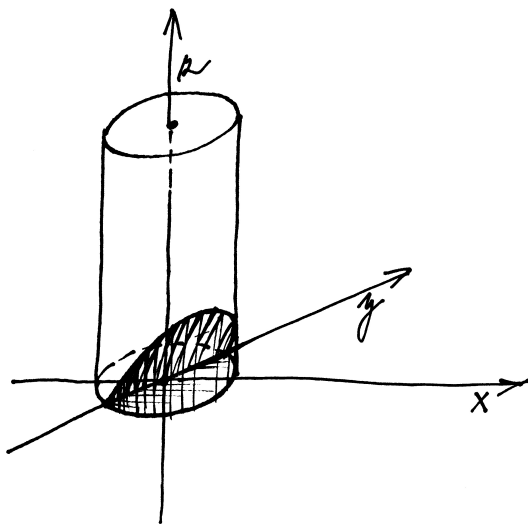
**8.H.24.** Find the maximum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = xy$  on the ellipse  $2x^2 + 3y^2 = 1$ .  $\circ$

**8.H.25.** Find the maximum of the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = xy$  on the ellipse  $x^2 + 2y^2 = 1$ .  $\circ$

**I. Volumes, areas, centroids of solids**

**8.I.1.** Find the volume of the solid which lies in the half-plane  $z \geq 0$ , the cylinder  $x^2 + y^2 \leq 1$ , and the half-plane

- a)  $z \leq x$ ,
- b)  $x + y + z \leq 0$ .



**Solution.** a) The volume can be calculated with ease using cylindrical coordinates. There, the cylinder is determined by the inequality  $r \leq 1$ ; the half-plane  $z \leq x$  by  $z \leq r \cos \varphi$ , then. Altogether, we get

$m + n = 2$  and to divide the matrix

$$(D^1 \tilde{F}^{-1}) = \begin{pmatrix} \text{id}_{\mathbb{R}^m} & 0 \\ D_x^1 F(x, y) & D_y^1 F(x, y) \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

into blocks of  $m$  and  $n$  rows and columns (for instance  $A$  is of type  $m \times m$ , while  $C$  is of type  $n \times m$ ). Now, the matrices  $A, B, C, D$  can be determined from the defining equality for the inverse:

$$\begin{pmatrix} \text{id}_{\mathbb{R}^m} & 0 \\ D_x^1 F(x, y) & D_y^1 F(x, y) \end{pmatrix} \cdot \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} \text{id}_{\mathbb{R}^m} & 0 \\ 0 & \text{id}_{\mathbb{R}^n} \end{pmatrix}.$$

Apparently, it follows that  $A = \text{id}_{\mathbb{R}^m}$ ,  $B = 0$ ,  $D = (D_y^1 F)^{-1}$ , and finally,  $D_x^1 F + D_y^1 F \cdot C = 0$ . The latter equality implies already the desired relation

$$D^1 G = C = -(D_y^1 F)^{-1} \cdot D_x^1 F.$$

This concludes the proof of the theorem.  $\square$

**8.1.26. The gradient of a function.** As seen in the previous paragraph, if  $F$  is a continuously differentiable function of  $n$  variables, the definition  $F(x_1, \dots, x_n) = b$  with a fixed value  $b \in \mathbb{R}$  defines the subset  $M_b \subset \mathbb{R}^n$  which mostly has the properties of an  $(n - 1)$ -dimensional hypersurface. To be more precise, if the vector of the partial derivatives



$$D^1 F = \left( \frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_n} \right)$$

is non-zero, the set  $M$  can be described locally as the graph of a continuously differentiable function of  $n - 1$  variables. In this connection, there are *level sets*  $M_b$ .

The vector  $D^1 F \in \mathbb{R}^n$  is called the *gradient of the function*  $F$ . In technical and physical literature, it is also often denoted as  $\text{grad } F$  or  $\nabla F$ .

Since  $M_b$  is given by a constant value of the function  $F$ , the derivatives of the curves lying in  $M$  have the property that the differential  $dF$  always evaluates to zero along them. For every such curve,  $F(c(t)) = b$ , hence

$$\frac{d}{dt} F(c(t)) = dF(c'(t)) = 0.$$

On the other hand, we can consider a general vector  $v = (v_1, \dots, v_n) \in \mathbb{R}^n$  and the magnitude of the corresponding directional derivative

$$|d_v F| = \left| \frac{\partial F}{\partial x_1} v_1 + \dots + \frac{\partial F}{\partial x_n} v_n \right| = \cos \varphi \|D^1 F\| \|v\|,$$

where  $\varphi$  is the angle between the directions of the vector  $v$  and the gradient  $D^1 F$ , see the discussion about angles of vectors and straight lines in the fourth chapter (cf. definition 4.1.15). Thus, the following result is observed:

Zde by se hodil obrazek, napr. obr. 2

$$V = \int_0^1 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{r \cos \varphi} r \, dz \, d\varphi \, dr = \frac{2}{3}.$$

b) We will reduce this problem to one that is completely analogous to the above part by rotating the solid around the  $z$ -axis by the angle  $\pi/4$  (be it in the positive or the negative direction). Applying the rotation matrix

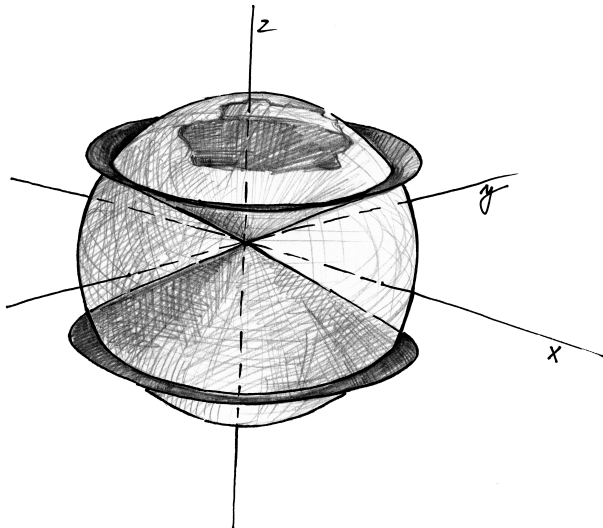
$$\begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

the original inequality  $x+y+z \leq 0$  is transformed to  $\sqrt{2}x' + z' \leq 0$  in the new coordinates. Now, it is easy to express the integral that corresponds to the volume of the examined solid:

$V = \int_0^1 \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_{-\sqrt{2}r \cos \varphi}^0 r \, dz \, d\varphi \, dr = \frac{2\sqrt{2}}{3}$ . We need not have computed the result as we did; instead, we could notice that the solid from part (a) differs only by homothety with coefficient  $\sqrt{2}$  in the direction of the  $y$ -axis. See also note 8.I.11.  $\square$

**8.I.2.** Find the volume of the solid in  $\mathbb{R}^3$  which is given by  $x^2 + y^2 + z^2 \leq 1, 3x^2 + 3y^2 \geq z^2, x \geq 0$ .

**Solution.**



First, we should realize what the examined solid looks like. It is a part of a ball which lies outside a given cone (see the picture).

The best way to determine the volume is probably to subtract half the volume of the sector given by the cone from half the ball's volume (note that the volume of the solid does not change if we replace the condition  $x \geq 0$  with  $z \geq 0$  – the sector is cut either “horizontally” or “vertically”, but always to halves). We will calculate in spherical coordinates.

$$\begin{aligned} x &= r \cos(\varphi) \sin(\psi), \\ y &= r \sin(\varphi) \sin(\psi), \\ z &= r \cos(\psi), \end{aligned}$$

$$\varphi \in [0, 2\pi), \psi \in [0, \pi), r \in (0, \infty).$$

THE MAXIMAL GROWTH OF A FUNCTION

**Proposition.** The gradient  $D^1F = \left( \frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_n} \right)$  provides the directions of maximal growth of the function  $F$  of  $n$  variables.

Moreover, the vanishing directional derivatives are exactly those in directions perpendicular to the gradient.

Therefore, it is clear that the tangent plane to a non-empty level set  $M_b$  in a neighbourhood of its point with non-zero gradient  $D^1F$  is determined by the orthogonal complement to the gradient, and the gradient itself is the *normal vector* of the hypersurface  $M_b$ .

For instance, considering a sphere in  $\mathbb{R}^3$  with radius  $r > 0$ , centered at  $(a, b, c)$ , i.e. given implicitly by the equation

$$F(x, y, z) = (x - a)^2 + (y - b)^2 + (z - c)^2 = r^2,$$

The normal vectors at a point  $P = (x_0, y_0, z_0)$  are obtained as non-zero multiples of the gradient, i.e. multiples of

$$D^1F = (2(x_0 - a), 2(y_0 - b), 2(z_0 - c)),$$

and the tangent vectors are exactly the vectors perpendicular to the gradient. Therefore, the tangent plane to a sphere at the point  $P$  can always be described implicitly in terms of the equation

$$0 = (x_0 - a)(x - x_0) + (y_0 - b)(y - y_0) + (z_0 - c)(z - z_0).$$

This is a special case of the following general formula:

TANGENT HYPERPLANES TO LEVEL SETS

**Theorem.** For a function  $F(x_1, \dots, x_n)$  of  $n$  variables and a point  $P = (p_1, \dots, p_n)$  in a level set  $M_b$  of the function  $F$  such that the gradient  $D^1F$  is non-vanishing at  $P$ , the implicit equation for the tangent hypersurface to  $M_b$  is

$$0 = \frac{\partial F}{\partial x_1}(P)(x_1 - p_1) + \dots + \frac{\partial F}{\partial x_n}(P)(x_n - p_n).$$

**PROOF.** The statement is clear from the previous discussions. The tangent hyperplane must be  $(n - 1)$ -dimensional, so its direction space is given as the kernel of the linear form given by the gradient (zero values of the corresponding linear mapping  $\mathbb{R}^n \rightarrow \mathbb{R}$  given by multiplying the column of coordinates by the row vector  $\text{grad } F$ ). Clearly, the selected point  $P$  satisfies the equation.  $\square$

**8.1.27. Illumination of 3D objects.** Consider the illumination of a three-dimensional object where the direction  $v$  of the light falling onto the two-dimensional surface  $M$  of this object is known. Assume  $M$  is given implicitly by an equation  $F(x, y, z) = 0$ .

The light intensity at a point  $P \in M$  is defined as  $I \cos \varphi$ , where  $\varphi$  is the angle between the normal line to  $M$  and the vector which is opposite to the flow of the light. As seen, the normal line is determined by the gradient of the function  $F$ .

The Jacobian of this transformation  $\mathbb{R}^3 \rightarrow \mathbb{R}^3$  is  $r^2 \sin(\psi)$ .

First of all, let us determine the volume of the ball. As for the integration bounds, it is convenient to express the conditions that bind the solid in the coordinates we will work in. In the spherical coordinates, the ball is given by the inequality

$$x^2 + y^2 + z^2 = r^2 \leq 1.$$

First, let us find the integration bounds for the variable  $\varphi$ . If we denote by  $\pi_\varphi$  the projection onto the  $\varphi$ -coordinate in the spherical coordinates ( $\pi_\varphi(\varphi, \theta, r) = \varphi$ ), then the image of the projection  $\pi_\varphi$  of the solid in question gives the integration bounds for the variable  $\varphi$ . We know that  $\pi_\varphi(\text{ball}) = [0, 2\pi)$  (the equation  $r^2 \leq 1$  does not contain the variable  $\varphi$ , so there are no constraints on it, and it takes on all possible values; this can also easily be imagined in space).

Having the bounds of one of the variables determined, we can proceed with the bounds of other variables. In general, those may depend on the variables whose bounds have already been determined (although this is not the case here). Thus, we choose arbitrarily a  $\varphi_0 \in [0, 2\pi)$ , and for this  $\varphi_0$  (fixed from now on), we find the intersection of the solid (ball) and the surface  $\varphi = \varphi_0$  and its projection  $\pi_\psi$  on the variable  $\psi$ . Similarly like for  $\varphi$ , the variable  $\psi$  is not bounded (either by the inequality  $r^2 \leq 1$  or the equality  $\varphi = \varphi_0$ ), so it can take on all possible values,  $\psi \in [0, \pi)$ .

Finally, let us fix a  $\varphi = \varphi_0$  and a  $\psi = \psi_0$ . Now, we are looking for the projection  $\pi_r(U)$  of the object (line segment)  $U$  given by the constraints  $r^2 \leq 1$ ,  $\varphi = \varphi_0$ ,  $\psi = \psi_0$  on the variable  $r$ . The only constraint for  $r$  is the condition  $r^2 \leq 1$ , so  $r \in (0, 1]$ .

Note that the integration bounds of the variables are independent of each other, so we can perform the integration in any order. Thus, we have

$$V_{\text{koule}} = \int_0^1 \int_0^{2\pi} \int_0^\pi r^2 \sin(\psi) \, d\psi \, d\varphi \, dr = \frac{4}{3}\pi.$$

Now, let us compute the volume of the spherical sector given by  $x^2 + y^2 + z^2 \leq 1$  and  $3x^2 + 3y^2 \geq z^2$ . Again, we express the conditions in the spherical coordinates:  $r^2 \leq 1$ ,  $3 \sin^2(\psi) \geq \cos^2(\psi)$ , i. e.,  $\text{tg}(\psi) \geq \frac{1}{\sqrt{3}}$ . Just like in the case of the ball, we can see that the variables occur independently in the inequalities, so the integration bounds of the variables will be independent of each other as well. The condition  $r^2 \leq 1$  implies  $r \in (0, 1]$ ; from  $\text{tg}(\psi) \geq \frac{1}{\sqrt{3}}$ , we have  $\psi \in [0, \frac{\pi}{6}]$ . The variable  $\varphi$  is not restricted by any condition, so  $\varphi \in [0, 2\pi]$ .

$$V_{\text{sector}} = \int_0^{2\pi} \int_0^1 \int_0^{\frac{\pi}{6}} r^2 \sin \psi \, d\psi \, dr \, d\varphi = \frac{2 - \sqrt{3}}{3}\pi,$$

altogether,

$$V = V_{\text{ball}} - V_{\text{sector}} = \frac{2}{3}\pi - \frac{2 - \sqrt{3}}{3}\pi = \frac{\pi}{\sqrt{3}}.$$

The sign of the expression then says which side of the surface is illuminated.

For example, consider an illumination with constant intensity  $I_0$  in the direction of the vector  $v = (1, 1, -1)$  (i.e. “downward askew”), and let the ball given by the equation  $F(x, y, z) = x^2 + y^2 + z^2 - 1 \leq 0$  be the object of interest. Then, for a point  $P = (x, y, z) \in M$  on the surface, the intensity

$$I(P) = \frac{\text{grad } F \cdot v}{\|\text{grad } F\| \|v\|} I_0 = \frac{-2x - 2y + 2z}{2\sqrt{3}} I_0$$

is obtained. Notice that, as anticipated, the point which is illuminated with the (full) intensity  $I_0$  is the point  $P = \frac{1}{\sqrt{3}}(-1, -1, 1)$  on the surface of the ball, while the antipodal point is fully illuminated with the minus sign (i.e. on the inside of the sphere).

Zde by se mozna hodilo i neco vic, aspon obrazek, neco jako obr. 3

**8.1.28. Tangent and normal spaces.** Ideas about tangent and normal lines can be extended to general dimensions. With a mapping  $F : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^n$ , and coordinate functions  $f_i$ , one can also consider the  $n$  equations for  $n + m$  variables

$$f_i(x_1, \dots, x_{m+n}) = b_i, \quad i = 1, \dots, n,$$

expressing the equality  $F(x) = b$  for a vector  $b \in \mathbb{R}^n$ .

Assuming that the conditions of the implicit function theorem hold, the set of all solutions  $(x_1, \dots, x_{m+n}) \in \mathbb{R}^{m+n}$  is (at least locally) the graph of a mapping  $G : \mathbb{R}^m \rightarrow \mathbb{R}^n$ . Technically, it is necessary to have some submatrix in  $D^1F$  of the maximal possible rank  $n$ .

For a fixed choice  $b = (b_1, \dots, b_n)$ , the set of all solutions is, of course, the intersection of all hypersurfaces  $M(b_i, f_i)$  corresponding to the particular functions  $f_i$ . The same must hold for tangent directions, while normal directions are generated by the individual gradients. Therefore, if  $D^1F$  is the Jacobi matrix of a mapping which implicitly defines a set  $M$  and  $P = (p_1, \dots, p_{m+n}) \in M$  is a point such that  $M$  is the graph of a mapping in the neighbourhood of the point  $P$ ,

$$D^1F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_{m+n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_{m+n}} \end{pmatrix},$$

then the affine subspace in  $\mathbb{R}^{m+n}$  which contains exactly all tangent lines going through the point  $P$  is given implicitly by the following equations:

$$0 = \frac{\partial f_1}{\partial x_1}(P)(x_1 - p_1) + \cdots + \frac{\partial f_1}{\partial x_n}(P)(x_{m+n} - p_{m+n})$$

⋮

$$0 = \frac{\partial f_n}{\partial x_1}(P)(x_1 - p_1) + \cdots + \frac{\partial f_n}{\partial x_n}(P)(x_{m+n} - p_{m+n}).$$

This subspace is called the *tangent space* to the (implicitly given)  $m$ -dimensional surface  $M$  at the point  $P$ .

The *normal space* at the point  $P$  is the affine subspace generated by the point  $P$  and the gradients of all the functions

We could also have computed the volume directly:

$$V = \int_0^\pi \int_0^1 \int_{\frac{\pi}{6}}^{\frac{5\pi}{6}} r^2 \sin \psi \, d\psi \, dr \, d\varphi = \frac{\pi}{\sqrt{3}}.$$

In cylindrical coordinates

$$x = r \cos(\varphi),$$

$$y = r \sin(\varphi),$$

$$z = z$$

with Jacobian  $r$  of this transformation, the calculation of the volume as the difference of the two solids considered above looks as follows:

$$V = \frac{2}{3}\pi - \int_0^{2\pi} \int_{\frac{1}{2}}^1 \int_0^1 r \, dz \, dr \, d\varphi = \frac{\pi}{\sqrt{3}}.$$

Note that we cannot compute the volume of the solid directly in the cylindrical coordinates. Thus, we must split it into two solids defined by the conditions  $r \leq \frac{1}{2}$  and  $r \geq \frac{1}{2}$ , respectively.

$$\begin{aligned} V &= V_1 + V_2 \\ &= \int_0^{2\pi} \int_0^{\frac{1}{2}} \int_0^{\sqrt{3}r} r \, dz \, dr \, d\varphi \\ &\quad + \int_0^{2\pi} \int_{\frac{1}{2}}^1 \int_0^{\sqrt{1-r^2}} r \, dz \, dr \, d\varphi \\ &= \frac{\pi}{\sqrt{3}}. \end{aligned}$$

□

Another alternative is to compute it as the volume of a solid of revolution, again splitting the solid into the two parts as in the previous case (the part “under the cone” and the part “under the sphere”). However, these solids cannot be obtained by rotating around one of the axes. The volume of the former part can be calculated as the difference between the volumes of the cylinder  $x^2 + y^2 \leq \frac{1}{4}$ ,  $0 \leq z \leq \frac{\sqrt{3}}{2}$  and the cone’s part  $3x^2 + 3y^2 \leq z^2$ ,  $0 \leq z \leq \frac{\sqrt{3}}{2}$ . The volume of the latter one is then the difference between the volumes of the solid that is created by rotating the part of the arc  $y = \sqrt{(1-x^2)}$ ,  $\frac{1}{2} \leq x \leq 1$  around the  $z$ -axis and the cylinder  $x^2 + y^2 \leq \frac{1}{4}$ ,  $0 \leq z \leq \frac{\sqrt{3}}{2}$ .

$$\begin{aligned} V &= V_1 + V_2 \\ &= \left( \frac{\pi\sqrt{3}}{8} - \frac{\pi\sqrt{3}}{24} \right) + \left( \pi \int_0^{\frac{\sqrt{3}}{2}} (1-r^2) \, dr - \frac{\pi\sqrt{3}}{8} \right) \\ &= \frac{\pi\sqrt{3}}{4} + \frac{\pi}{4\sqrt{3}} = \frac{\pi}{\sqrt{3}}. \end{aligned}$$

**8.I.3.** Calculate the volume of the spherical segment of the ball  $x^2 + y^2 + z^2 = 2$  cut by the plane  $z = 1$ .

$f_1, \dots, f_n$  at the point  $P$ , i.e. the rows of the Jacobi matrix  $D^1 F$ .

As an illustrative simple example, calculate the tangent and normal spaces to a conic section in  $\mathbb{R}^3$ . Consider the equation of a cone with vertex at the origin,

$$0 = f(x, y, z) = z - \sqrt{x^2 + y^2},$$

and a plane, given by

$$0 = g(x, y, z) = z - 2x + y + 1.$$

The point  $P = (1, 0, 1)$  belongs to both the cone and the plane, so the intersection  $M$  of these surfaces is a curve (draw a diagram!). Its tangent line at the point  $P$  is given by the following equations:

$$\begin{aligned} 0 &= -\frac{1}{2\sqrt{x^2 + y^2}} 2x \Big|_{x=1, y=0} \cdot (x-1) \\ &\quad - \frac{1}{2\sqrt{x^2 + y^2}} 2y \Big|_{x=1, y=0} \cdot y + 1 \cdot (z-1) \\ &= -x + z \end{aligned}$$

$$0 = -2(x-1) + y + (z-1) = -2x + y + z + 1,$$

while the plane perpendicular to the curve, containing the point  $P$ , is given parametrically by the expression

$$(1, 0, 1) + \tau(-1, 0, 1) + \sigma(-2, 1, 1)$$

with real parameters  $\tau$  and  $\sigma$ .

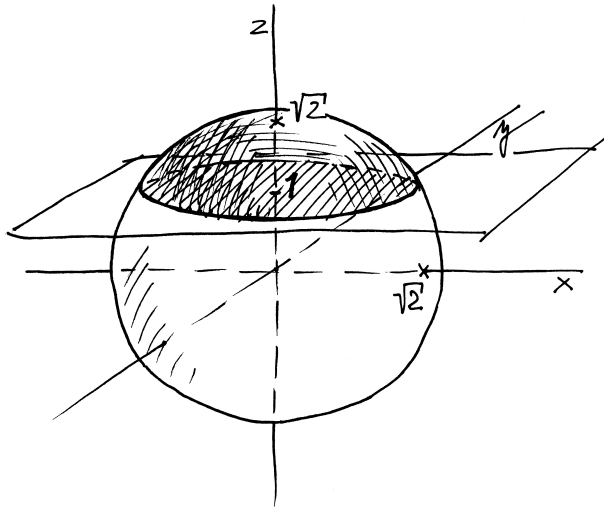
**8.1.29. Constrained extrema.** Now we come with the first really serious application of the differential calculus of more variables. The typical task in optimization is to find the extrema of values depending on several (yet finitely many) parameters, under some further constraints on the parameters of the model.

The problem often has  $m+n$  parameters constrained by  $n$  conditions. In the language of differential calculus, it is desired to find the extrema of a differentiable function  $h$  on the set  $M$  of points given implicitly by a vector equation  $F(x_1, \dots, x_{m+n}) = 0$ . Of course, we might first locally parameterize the solution space of the latter equation by  $m$  free parameters, express the function  $h$  in terms of these parameters and look for the local extrema by inspecting the critical points. However, we have already prepared more efficient procedures for this effort.

For every curve  $c(t) \subset M$  going through  $P = c(0)$ , it must be ensured that  $h(c(t))$  is an extremum of this univariate function. Therefore, the derivative must satisfy

$$\frac{d}{dt} h(c(t))|_{t=0} = d_{c'(0)} h(P) = dh(P)(c'(0)) = 0.$$

This means that the differential of the function  $h$  at the point  $P$  is zero along all tangent increments to  $M$  at  $P$ . This property is equivalent to stating that the gradient of  $h$  lies in the normal subspace (more precisely, in the modelling vector space of the normal subspace). Such points  $P \in M$  are called *stationary points* of the function  $h$  with respect to the constraints given by  $F$ .



**Solution.** We will compute the integral in spherical coordinates. The segment can be perceived as a spherical sector without the cone (with vertex at the point  $[0, 0, 0]$  and the circular base  $z = 1, x^2 + y^2 = 1$ ). In these coordinates, the sector is the product of the intervals  $[0, \sqrt{2}] \times [0, 2\pi] \times [0, \pi/4]$ . We thus integrate in the given bounds, in any order:

$$\int_0^{2\pi} \int_0^{\sqrt{2}} \int_0^{\pi/4} r^2 \sin(\theta) \, d\theta \, dr \, d\varphi = \frac{4}{3}(\sqrt{2} - 1)\pi.$$

In the end, we must subtract the volume of the cone. That is equal to  $\frac{1}{3}\pi R^2 H$  (where  $R$  is the radius of the cone's base and  $H$  is its height; both are equal to 1 in our case), so the total volume is

$$V_{\text{sector}} - V_{\text{cone}} = \frac{4}{3}(\sqrt{2} - 1)\pi - \frac{1}{3}\pi = \frac{1}{3}\pi(4\sqrt{2} - 5).$$

The volume of a general spherical segment with height  $h$  in a ball with radius  $R$  could be computed similarly:

$$\begin{aligned} V &= V_{\text{sector}} - V_{\text{cone}} \\ &= \int_0^{2\pi} \int_0^{\arccos\left(\frac{R-h}{R}\right)} \int_0^R r^2 \sin(\theta) \, dr \, d\theta \, d\varphi \\ &\quad - \frac{1}{3}\pi(2Rh - h^2)(R - h) \\ &= \frac{1}{3}\pi h^2(3R - h). \end{aligned}$$

**8.I.4.** Find the volume of the part of the cylinder  $x^2 + z^2 = 16$  which lies inside the cylinder  $x^2 + y^2 = 16$ .

As seen in the previous paragraph, the normal space to the set  $M$  is generated by the rows of the Jacobi matrix of the mapping  $F$ , so the stationary points are described equivalently by the following proposition:

LAGRANGE MULTIPLIERS

**Theorem.** Let  $F = (f_1, \dots, f_n) : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^n$  be a differentiable function in a neighbourhood of a point  $P$ ,  $F(P) = 0$ . Further, let  $M$  be given implicitly by an equation  $F(x, y) = 0$ , and let the rank of the matrix  $D^1F$  at the point  $P$  be  $n$ . Then  $P$  is a stationary point of a continuously differentiable function  $h : \mathbb{R}^{m+n} \rightarrow \mathbb{R}$  with respect to the constraints  $F$ , if and only if there exist real parameters  $\lambda_1, \dots, \lambda_n$  such that

$$\text{grad } h = \lambda_1 \text{grad } f_1 + \dots + \lambda_n \text{grad } f_n.$$

The procedure suggested by the theorem is called the *method of Lagrange multipliers*. It is of algorithmic character. Consider the numbers of unknowns and equations: the gradients are vectors of  $m + n$  coordinates, so the statement of the theorem yields  $m + n$  equations. The variables are, on one side, the coordinates  $x_1, \dots, x_{m+n}$  of the stationary points  $P$  with respect to the constraints, and, on the other hand, the  $n$  parameters  $\lambda_i$  in the linear combination. It remains to say that the point  $P$  belongs to the implicitly given set  $M$ , which represents  $n$  further equations. Altogether, there are  $2n + m$  equations for  $2n + m$  variables, so it can be expected that the solution is given by a discrete set of points  $P$  (i.e., each one of them is an isolated point).

Very often, the system of equations is a seemingly simple system of algebraic equations, but in fact only rarely can it be solved explicitly. We return to special algebraic methods for systems of polynomial equations in chapter 12. There are also various numerical approaches to such systems. Theoretical details are not discussed here, but there are several solved examples in the other column, including also the illustration of how to use the second derivatives to decide about the local extrema under the constraints.

**8.1.30. Arithmetic mean versus geometric mean.** As an example of practical application of the Lagrange multipliers, we prove the inequality

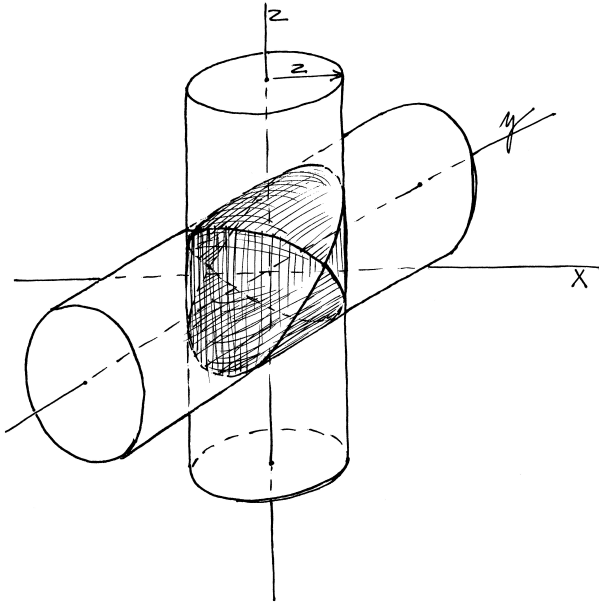
$$\frac{1}{n}(x_1 + \dots + x_n) \geq \sqrt[n]{x_1 \cdots x_n}$$

for any  $n$  positive real numbers  $x_1, \dots, x_n$ . Equality occurs if and only if all the  $x_i$ 's are equal.

Consider the sum  $x_1 + \dots + x_n = c$  as the constraint for a (non-specified) non-negative constant  $c$ . We look for the maxima and minima of the function

$$f(x_1, \dots, x_n) = \sqrt[n]{x_1 \cdots x_n}$$

with respect to the constraint and the assumption  $x_1 > 0, \dots, x_n > 0$ .



The normal vector to the part of the hyperplane  $M$  defined by the constraints is  $(1, \dots, 1)$ . Therefore, the function  $f$  can have an extremum only at those points where its gradient is a multiple of this normal vector. Hence there is the following system of equations for the desired points (the components  $\frac{\partial f}{\partial x_i}$  of the gradient appear on the left-hand side,  $\lambda$ -multiple of the normal vector is on the right):

$$\frac{1}{n} \frac{1}{x_i} \sqrt[n]{x_1 \cdots x_n} = \lambda,$$

for  $i = 1, \dots, n$  and  $\lambda \in \mathbb{R}$ .

These equations imply  $x_1 = \dots = x_n$  in the set  $M$ . If the variables  $x_i$  are allowed to be zero as well, then the set  $M$  would be compact, so the function  $f$  would have to have both a maximum and a minimum there. However,  $f$  is minimal if and only if at least one of the values  $x_i$  is zero; so the function necessarily has a strict maximum at the point with  $x_i = \frac{c}{n}$ ,  $i = 1, \dots, n$ , and then  $\lambda = \frac{1}{n}$ .

By substituting, the geometric mean equals the arithmetic mean for these extreme values, but it is strictly smaller at all other points with the given sum  $c$  of coordinates, which proves the inequality.

**Solution.** We will compute the integral in Cartesian coordinates. Since the solid is symmetric, it suffices to integrate over the first octant (interchanging  $x$  and  $-x$  does not change the equation of the solid; the same holds for  $y$  and for  $z$ ). The part of the solid that lies in the first octant is given by the space under the graph of the function  $z(x, y) = \sqrt{16 - x^2}$  and over the quarter-disc  $x^2 + y^2 \leq 16$ ,  $x \geq 0$ ,  $y \geq 0$ . Therefore, the volume of the whole solid is equal to

$$V = 8 \int_0^4 \int_0^{\sqrt{16-x^2}} \frac{4}{\sqrt{16-x^2}} dy dx = 128. \quad \square$$

**Remark.** Note that the projection of the considered solid onto both the plane  $y = 0$  and the plane  $z = 0$  is a circle with radius 4, yet the solid is not a ball.

## 2. Integration for the second time

We return to the process of integration, discussed in the second and third parts of chapter six. We saw that the integration with respect to the diverse coordinates can be iterated. Now we extend the concept of the Riemann integration and Jordan measure to general Euclidean spaces and, again, we shall see that the approaches coincide for many reasonable functions.

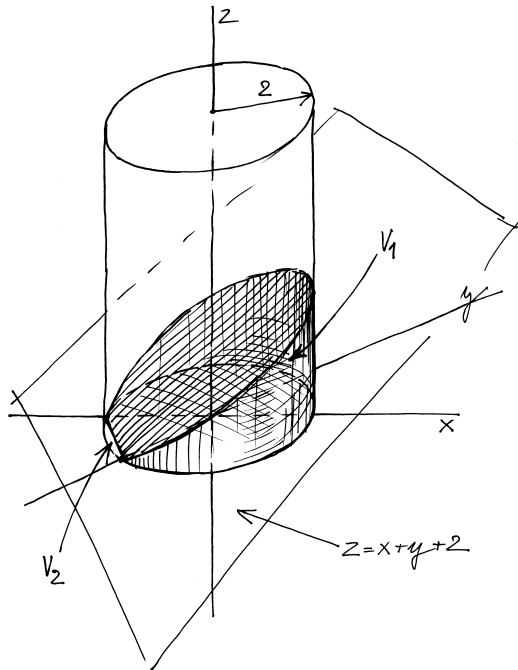
### 8.2.1. Integrals dependent on parameters.



Recall that integrating a function  $f(x, y_1, \dots, y_n)$  of  $n + 1$  variables with respect to the single variable  $x$ , the result is a function  $F(y_1, \dots, y_n)$  of the remaining variables. Essentially, we proved the following theorem already in 6.3.11 and 6.3.13. This is an extremely useful technical tool, as we saw when handling the Fourier transforms and convolutions in the last chapter. Previous results about extrema of multivariate functions also have a direct application for minimization of areas or volumes of objects defined in terms of functions dependent on parameters, etc.



**8.I.5.** Find the volume of the part of the cylinder  $x^2 + y^2 = 4$  bounded by the planes  $z = 0$  and  $z = x + y + 2$ .



**Solution.** We will work in cylindric coordinates given by the equations  $x = r \cos(\varphi)$ ,  $y = r \sin(\varphi)$ ,  $z = z$ . The Jacobian of this transformation is  $J = r$ . The solid can be divided into two parts: above and below the plane  $z = 0$ , whose volumes will be denoted by  $V_1$  and  $V_2$ , respectively. Further, we can notice that one part of the solid with volume  $V_1$  is a pyramid with vertices  $[0, 0, 0]$ ,  $[0, 0, 2]$ ,  $[-2, 0, 0]$ ,  $[0, -2, 0]$ . Thus, we will further split this solid (above  $z = 0$ ) into two parts, whose volumes we will calculate separately.

$$\begin{aligned}
 V_1 - V_{\text{pyramid}} &= \int_{-\pi/2}^{\pi} \left( \int_0^2 [r \sin \varphi + r \cos \varphi + 2] r \, dr \right) d\varphi, \\
 &= 6\pi + \frac{16}{3}, \\
 V_{\text{pyramid}} &= \frac{4}{3}
 \end{aligned}$$

Further,

$$V_1 - V_2 = \int_{-\pi}^{\pi} \int_0^2 r^2 (\sin(\varphi) + \cos(\varphi)) + 2r \, dr \, d\varphi = 8\pi,$$

so  $V_1 + V_2 = 4\pi + \frac{40}{3}$ .  $\square$

**Remark.** During the calculation, we made use of the fact that integrating a function of two variables over an area in  $\mathbb{R}^2$  yields the difference of the volume of the solid in  $\mathbb{R}^3$  determined by the graph of the integrated function and lying above  $z = 0$  and the one lying below  $z = 0$ .

**8.I.6.** Find the volume of the solid in  $\mathbb{R}^3$  which is given by the intersection of the sphere  $x^2 + y^2 + z^2 = 4$  and the cylinder  $x^2 + y^2 = 1$ .

CONTINUITY AND DIFFERENTIATION

**Theorem.** Consider a continuous function  $f(x, y_1, \dots, y_n)$  defined for all  $x$  from a finite interval  $[a, b]$  and for all  $(y_1, \dots, y_n)$  lying in some neighbourhood  $U$  of a point  $c = (c_1, \dots, c_n) \in \mathbb{R}^n$ , and its integral

$$F(y_1, \dots, y_n) = \int_a^b f(x, y_1, \dots, y_n) \, dx.$$

Then function  $F(y_1, \dots, y_n)$  is continuous on  $U$ .

Moreover, if there exists the continuous partial derivative  $\frac{\partial f}{\partial y_j}$  on a neighbourhood of the point  $c$ , then  $\frac{\partial F}{\partial y_j}(c)$  exists as well and

$$\frac{\partial F}{\partial y_j}(c) = \int_a^b \frac{\partial f}{\partial y_j}(x, c_1, \dots, c_n) \, dx.$$

**PROOF.** In Chapter 6, we dealt with two variables  $x, y$  only, but replacing the absolute value  $|y|$  with the norm  $\|y\|$  of the vector of parameters does not change the argumentation at all. Again, the main point is that the continuous real functions on compact sets are uniformly continuous.

Since partial derivative concerns only one of the variables, the rest of the theorem was proved in 6.3.13, too.  $\square$

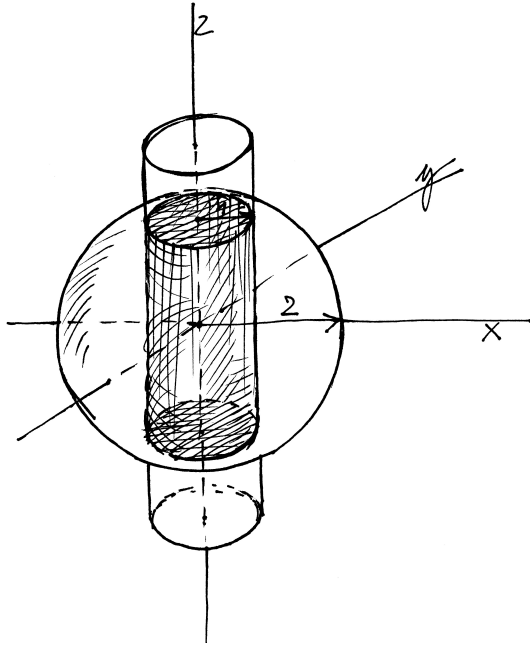
**8.2.2. Integration of multivariate functions.** In the case of univariate functions, integration is motivated by the idea of the area under the graph of a given function of one variable. Consider now the volume of the part of the three-dimensional space which lies under the graph of a function  $z = f(x, y)$  of two variables, and the multidimensional analogues in general.

In chapter six, small intervals  $[x_i, x_{i+1}]$  were chosen of length  $\Delta x_i$  which divided the whole interval  $[a, b]$ . Then, their representatives  $\xi_i$  were selected, and the corresponding part of the area was approximated by the area of the rectangle with height given by the value  $f(\xi_i)$  at the representative, i.e. the expression  $f(\xi_i)\Delta x_i$ .

In the case of functions of two variables, work with divisions in both variables and the values representing the height of the graph above the particular little rectangles in the plane.

The first thing to deal with is to determine the integration domain, that is, the region the function  $f$  is to be integrated over. As an example, consider the function  $z = f(x, y) = \sqrt{1 - x^2 - y^2}$ , whose graph is, inside the unit disc, half of the unit sphere. Integrating this function over the unit disc yields the volume of the unit semi-ball.

The simplest approach is to consider only those integration domains  $M$  which are given by products of intervals, i.e. given by ranges  $x \in [a, b]$  and  $y \in [c, d]$ . In this context, it is called a *multidimensional interval*. If  $M$  is a different bounded set in  $\mathbb{R}^2$ , work with a sufficiently large area  $[a, b] \times [c, d]$ , rather than with the set itself, and adjust the function so that  $f(x, y) = 0$  for all points lying outside  $M$ . Considering the above case of the unit ball, integrate over the

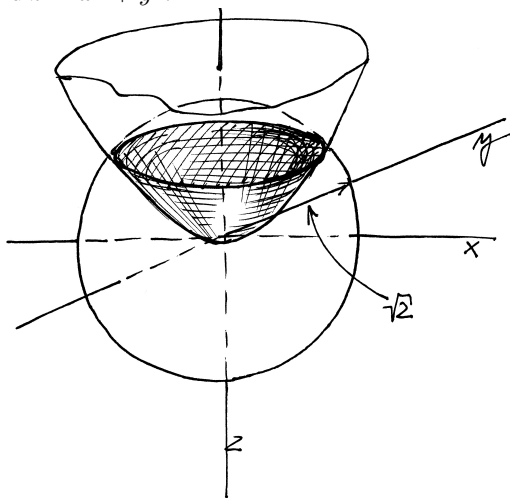


**Solution.** Thanks to symmetry, it suffices to compute the volume of the part that lies in the first octant. We will integrate in cylindrical coordinates given by the equations  $x = r \cos(\varphi)$ ,  $y = r \sin(\varphi)$ ,  $z = z$  with Jacobian  $J = r$ , and it is the space between the plane  $z = 0$  and the graph of the function  $z = \sqrt{4 - x^2 - y^2} = \sqrt{4 - r^2}$ . Therefore, we can directly write is as the double integral

$$V = 8 \int_0^{\pi/2} \int_0^1 r \sqrt{4 - r^2} dr d\varphi = \frac{2}{3} (8 - 3\sqrt{3})\pi.$$

□

**8.I.7.** Find the volume of the solid in  $\mathbb{R}^3$  which is given by the intersection of the sphere  $x^2 + y^2 + z^2 = 2$  and the paraboloid  $z = x^2 + y^2$ .



**Solution.** Once again, we will work in cylindrical coordinates:

$$V = \int_0^{2\pi} \int_0^1 \int_{r^2}^{\sqrt{2-r^2}} r dz dr d\varphi = \frac{4\sqrt{2}\pi}{3} - \frac{7\pi}{6}.$$

set  $M = [-1, 1] \times [-1, 1]$  the function

$$f(x, y) = \begin{cases} \sqrt{1 - x^2 - y^2} & \text{for } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The definition of the Riemann integral then faithfully follows the procedure from paragraph 6.2.8. This can be done for an arbitrary finite number of variables.

Given an  $n$ -dimensional interval  $I$  and partitions into  $k_i$  subintervals in each variable  $x_i$ , select the partition of  $I$  into  $k_1 \cdots k_n$  small  $n$ -dimensional intervals, and write  $\Delta x_{i_1 \dots i_n}$  for their volumes. The maximum of the lengths of the sides of the multidimensional intervals in such a partition is called its *norm*.

RIEMANN INTEGRAL

**Definition.** The Riemann integral of a real-valued function  $f$  defined on a multidimensional interval  $I = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n]$  exists if for every choice of a sequence of divisions  $\Xi$  (dividing the multidimensional interval in all variables simultaneously), and the representatives  $\xi_{i_1 \dots i_n}$  of the little multidimensional intervals in the partitions, with the norm of the partitions converging to zero, the integral sums

$$S_{\Xi, \xi} = \sum_{i_1 \dots i_n} f(\xi_{i_1 \dots i_n}) \Delta x_{i_1 \dots i_n}$$

always converge to the value

$$S = \int_I f(x_1, \dots, x_n) dx_1 \dots dx_n,$$

independent of the selected sequence of divisions and representatives.

The function  $f$  is then said to be Riemann-integrable over  $I$ .

As a relatively simple exercise, prove in detail that every Riemann-integrable function over an interval  $I$  must be bounded there. The reason is the same as in the case of univariate functions: the control of the norms of the divisions used in the definition is somewhat rough.

The situation gets worse when integrating in this way over unbounded intervals, see more remarks in 8.2.6 below. Therefore, consider integration of functions over  $\mathbb{R}^n$  mainly for functions whose support is compact, that is, functions which vanish outside a bounded interval  $I$ .

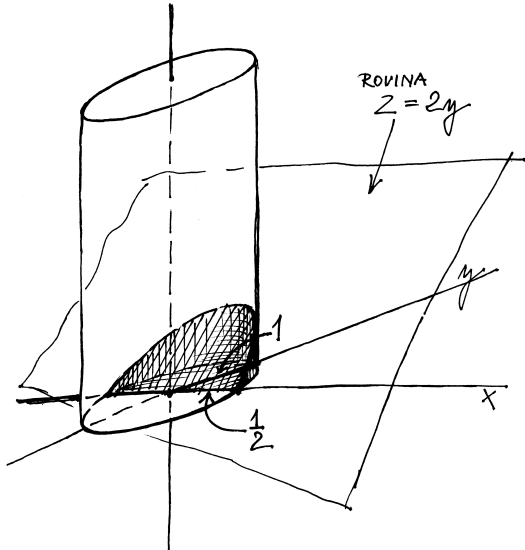
A bounded set  $M \subset \mathbb{R}^n$  is said to be *Riemann-measurable*<sup>4</sup> if and only if its indicator function, defined by

$$\chi_M(x_1, \dots, x_n) = \begin{cases} 1 & \text{for } (x_1, \dots, x_n) \in M \\ 0 & \text{for all other points in } \mathbb{R}^n, \end{cases}$$

is Riemann-integrable over  $\mathbb{R}^n$ .

<sup>4</sup> Better to say “measurable via Riemann integration”, the measure itself is commonly called the *Peano–Jordan measure* in the literature.

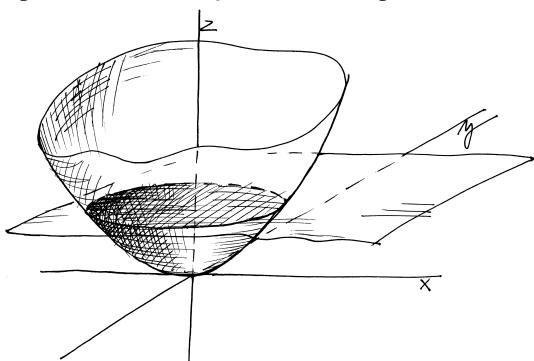
**8.I.8.** Find the volume of the solid in  $\mathbb{R}^3$  which is bounded by the elliptic cylinder  $4x^2 + y^2 = 1$  and the planes  $z = 2y$  and  $z = 0$ , lying above the plane  $z = 0$ .



**Solution.** Thanks to symmetry, it is advantageous to work in the coordinates  $x = \frac{1}{2}r \cos(\varphi)$ ,  $y = r \sin(\varphi)$ ,  $z = z$  with Jacobian  $J = \frac{1}{2}r$ . The equation of the elliptic cylinder in these coordinates is  $r^2 = 1$ . Thus, the wanted volume is

$$\begin{aligned} V &= \int_0^\pi \int_0^1 r \sin(\varphi) \frac{1}{2}r \, dr \, d\varphi \\ &= \int_0^\pi \int_0^1 r^2 \sin(\varphi) \, dr \, d\varphi = \int_0^\pi \frac{1}{3} \sin(\varphi) \, d\varphi = \frac{2}{3}. \end{aligned}$$

**8.I.9.** Find the volume of the solid in  $\mathbb{R}^3$  which is bounded by the paraboloid  $2x^2 + y^2 = z$  and the plane  $z = 2$ .



**Solution.** Similarly to the above problem, we choose “special” coordinates which respect the symmetry of the solid:  $x = \frac{1}{\sqrt{2}}r \cos(\varphi)$ ,  $y = r \sin(\varphi)$ ,  $z = z$  with Jacobian

□ For any Riemann-measurable set  $M$  and a function  $f$  defined at all points of  $M$ , consider the function  $\tilde{f} = \chi_M \cdot f$  as a function defined on the whole  $\mathbb{R}^n$ . This function  $\tilde{f}$  apparently has a compact support. The Riemann integral of the function  $f$  over the set  $M$  is defined by

$$\int_M f \, dx_1 \dots dx_n = \int_{\mathbb{R}^n} \tilde{f} \, dx_1 \dots dx_n,$$

supposing the integral on the right-hand side exists.

**8.2.3. Properties of Riemann integral.** This definition of the integral does not provide reasonable instructions for computing the values of Riemann integrals. However, it does lead to the following basic properties of the Riemann integral (cf. Theorem 6.2.8):

**Theorem.** The set of Riemann-integrable real-valued functions over a Riemann measurable domain  $M \subset \mathbb{R}^n$  is a vector space over the real scalars, and the Riemann integral is a linear form there.

If the integration domain  $M$  is given as a disjoint union of finitely many Riemann-measurable domains  $M_i$ , then  $f$  is integrable on  $M$  if and only if it is integrable on all  $M_i$ , and the integral over a function  $f$  over  $M$  is given by the sum of the integrals over the individual subdomains  $M_i$ .

**PROOF.** All the properties follows directly from the definition of the Riemann integral and the properties of convergent sequences of real numbers, just as in the case of univariate functions. Think out the details by yourselves. □

For practical use, rewrite the theorem into the usual equalities:

FINITE ADDITIVITY AND LINEARITY

Any linear combination of Riemann-integrable functions  $f_i : \mathbb{I} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, k$  (over scalars in  $\mathbb{R}$ ) is again a Riemann-integrable function, and its integral can be computed as follows:

$$\begin{aligned} &\int_I (a_1 f_1(x_1, \dots, x_n) + \dots + a_k f_k(x_1, \dots, x_n)) \, dx_1 \dots dx_n \\ &= a_1 \int_I f_1(x_1, \dots, x_n) \, dx_1 \dots dx_n + \\ &\quad \dots + a_k \int_I f_k(x_1, \dots, x_n) \, dx_1 \dots dx_n. \end{aligned}$$

Let  $M_1$  and  $M_2$  be disjoint Riemann-measurable sets, consider a function  $f : M_1 \cup M_2 \rightarrow \mathbb{R}$ . Then  $f$  is Riemann-integrable over both sets  $M_i$  if and only if it is integrable over its union, and

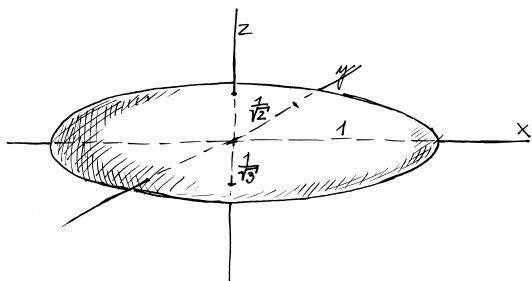
$$\begin{aligned} &\int_{M_1 \cup M_2} f(x_1, \dots, x_n) \, dx_1 \dots dx_n \\ &= \int_{M_1} f(x_1, \dots, x_n) \, dx_1 \dots dx_n + \\ &\quad \int_{M_2} f(x_1, \dots, x_n) \, dx_1 \dots dx_n. \end{aligned}$$

$J = \frac{1}{\sqrt{2}}r$ . The equation of the paraboloid in these coordinates is  $z = r^2$ , so the volume of the solid is equal to

$$\begin{aligned} V &= 4 \int_0^{\pi/2} \int_0^{\sqrt{2}} \int_{r^2}^2 \frac{1}{\sqrt{2}}r \, dz \, dr \, d\varphi \\ &= 2\sqrt{2} \int_0^{\pi/2} \int_0^{\sqrt{2}} 2r - r^3 \, dr \, d\varphi = 2\sqrt{2} \int_0^{\pi/2} d\varphi \\ &= \sqrt{2}\pi. \end{aligned}$$

□

**8.I.10.** Calculate the volume of the ellipsoid  $x^2 + 2y^2 + 3z^2 = 1$ .



**Solution.** We will consider the coordinates

$$\begin{aligned} x &= r \cos(\varphi) \sin(\theta), \\ y &= \frac{1}{\sqrt{2}}r \sin(\varphi) \sin(\theta), \\ z &= \frac{1}{\sqrt{3}}r \cos(\theta). \end{aligned}$$

The corresponding Jacobian is  $\frac{1}{\sqrt{6}}r^2 \sin(\theta)$ , so the volume is

$$V = \int_0^{2\pi} \int_0^{\pi} \int_0^1 \frac{1}{\sqrt{6}}r^2 \sin(\theta) \, dr \, d\theta \, d\varphi = \frac{4}{3\sqrt{6}}\pi.$$

□

**8.I.11. Remark.** Note that if the transformation the coordinates is linear (and affine), then the space is deformed “uniformly”. This means that the volume of an arbitrary solid is changed proportionally to the change of the volume of an infinitesimal volume element, which is the Jacobian. Therefore, if we consider the volume of the ball with a given radius  $r$  to be known, (in this case,  $r = 1$ ), we can infer directly that the volume of the ellipsoid is  $V = \frac{1}{\sqrt{6}} \cdot \frac{4}{3}\pi = \frac{4}{3\sqrt{6}}\pi$ .

**8.I.12.** Find the volume of the solid which is bounded by the paraboloid  $2x^2 + 5y^2 = z$  and the plane  $z = 1$ .

**Solution.** We choose the coordinates

$$\begin{aligned} x &= \frac{1}{\sqrt{2}}r \cos(\varphi), \\ y &= \frac{1}{\sqrt{5}}r \sin(\varphi), \\ z &= z. \end{aligned}$$

The determinant of the Jacobian is  $\frac{r}{\sqrt{10}}$ , so the volume is

$$V = \int_0^{2\pi} \int_0^1 \int_{r^2}^1 \frac{r}{\sqrt{10}} \, dz \, dr \, d\varphi = \frac{\pi}{2\sqrt{10}}.$$

□

**8.2.4. Multiple integrals.** Riemann-measurable domains especially involve cases when the boundary of the integration domain  $M$  can be expressed step by step via continuous dependencies between the coordinates in the following way. The first coordinate  $x$  runs within an interval  $[a, b]$ . The interval range of the next coordinate can be defined by two functions, i.e.  $y \in [\varphi(x), \psi(x)]$ , then the range of the next coordinate is expressed as  $z \in [\eta(x, y), \zeta(x, y)]$ , and so on for all of the other coordinates.



For example, this is easy in the case of a ball from the introductory example: for  $x \in [-1, 1]$ , define the range for  $y$  as  $y \in [-\sqrt{1-x^2}, \sqrt{1-x^2}]$ . The volume of the ball can then be computed by integration of the mentioned function  $f$ , or integrate the indicator function of the ball, i.e. the function which takes one on the subset  $M \subset \mathbb{R}^3$  which is further defined by  $z \in [-\sqrt{1-x^2-y^2}, \sqrt{1-x^2-y^2}]$ .

The following fundamental theorem transforms the computation of a Riemann integral to a sequence of computations of univariate integrals (while the other variables are considered to be parameters, which can appear in the integration bounds as well). Notice, we could have defined the multiple integral directly via the one-dimensional integration, but we would face the trouble of ensuring the independence of the result on our way of describing  $M$ . The theorem reveals that the two approaches coincide and there are no unclear points left.

#### MULTIPLE INTEGRALS

**Theorem.** Let  $M \subset \mathbb{R}^n$  be a bounded set, expressed with the help of continuous functions  $\psi_i, \eta_i$

$$M = \{(x_1, \dots, x_n); x_1 \in [a, b], x_2 \in [\psi_2(x_1), \eta_2(x_1)], \dots, x_n \in [\psi_n(x_1, \dots, x_{n-1}), \eta_n(x_1, \dots, x_{n-1})]\},$$

and let  $f$  be a function which is continuous on  $M$ . Then the Riemann integral of the function  $f$  over the set  $M$  exists and is given by the formula

$$\int_M f(x_1, x_2, \dots, x_n) \, dx_1 \dots dx_n = \int_a^b \left( \int_{\psi_2(x_1)}^{\eta_2(x_1)} \dots \left( \int_{\psi_n(x_1, \dots, x_{n-1})}^{\eta_n(x_1, \dots, x_{n-1})} f(x_1, x_2, \dots, x_n) \, dx_n \right) \dots dx_2 \right) dx_1$$

where the individual integrals are the one-variable Riemann integrals.



**PROOF.** Consider first the proof for the case of two variables. It can then be seen that there is no need of further ideas in the general case.

Consider an interval  $I = [a, b] \times [c, d]$  containing the set  $M = \{(x, y); x \in [a, b], y \in [\psi(x), \eta(y)]\}$  and divisions  $\Xi$  of the interval  $I$  with representatives  $\xi_{ij}$ .

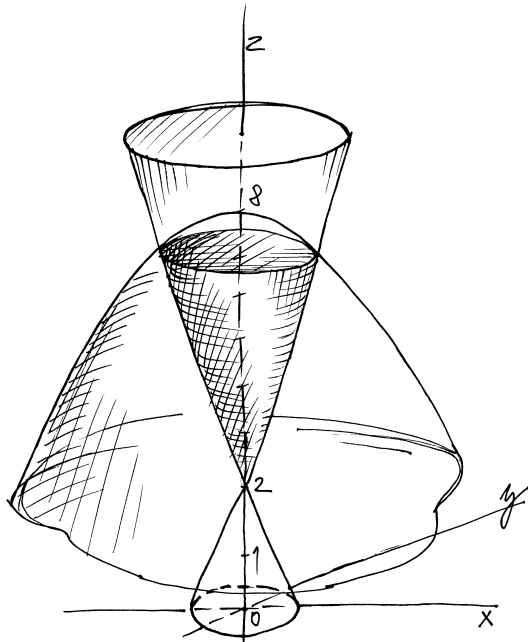
**8.I.13.** Find the volume of the solid which lies in the first octant and is bounded by the surfaces  $y^2 + z^2 = 9$  and  $y^2 = 3x$ .

**Solution.** In cylindrical coordinates,

$$V = \int_0^{\pi/2} \int_0^3 \int_0^{\sqrt{\frac{x}{3}} \cos^2(\varphi)} r \, dx \, dr \, d\varphi = \frac{27}{16}\pi.$$

□

**8.I.14.** Find the volume of the solid in  $\mathbb{R}^3$  which is bounded by the cone part  $2x^2 + y^2 = (z-2)^2$ ,  $z \geq 2$  and the paraboloid  $2x^2 + y^2 = 8 - z$ .



**Solution.** First of all, we find the intersection of the given surfaces:

$$(z - 2)^2 = -z + 8, \quad z \geq 2;$$

therefore,  $z = 4$ , and the equation of the intersection is  $2x^2 + y^2 = 4$ . The substitution  $x = \frac{1}{\sqrt{2}}r \cos(\varphi)$ ,  $y = r \sin(\varphi)$ ,  $z = z$  transforms the given surfaces to the form  $r^2 = (z-2)^2$ ,  $z \geq 2$ , and  $r^2 = 8 - z$ , i. e.,  $z = r + 2$  for the former surface and  $z = 8 - r^2$  for the latter. Altogether, the projection of the given solid onto the coordinate  $\varphi$  is equal to the interval  $[0, 2\pi]$ . Having fixed a  $\varphi_0 \in [0, 2\pi]$ , the projection of the intersection of the solid and the plane  $\varphi = \varphi_0$  onto the coordinate  $r$  equals (independently of  $\varphi_0$ ) the interval  $[0, 2]$ . Having fixed both  $r_0$  and  $\varphi_0$ , the projection of the intersection of the solid and the line  $r = r_0$ ,  $\varphi = \varphi_0$ , onto the coordinate  $z$  is equal to the interval  $[r_0 + 2, 8 - r_0^2]$ . The Jacobian of the considered transformation is  $J = \frac{1}{\sqrt{2}}r$ , so we can write

$$V = \int_0^{2\pi} \int_0^2 \int_{r+2}^{8-r^2} \frac{r}{\sqrt{2}} \, dz \, dr \, d\varphi = \frac{16\sqrt{2}}{3}\pi.$$

The corresponding integral sum is

$$\begin{aligned} S_{\Xi, \xi} &= \sum_{i,j} f(\xi_{ij}) \Delta x_{ij} \\ &= \sum_i \left( \sum_j f(\xi_{ij}) \Delta y_j \right) \Delta x_i, \end{aligned}$$

where  $\Delta x_{ij}$  is written for the product of the sizes  $\Delta x_i$  and  $\Delta x_j$  of the intervals which correspond to the choice of the representative  $\xi_{ij}$ .

Assume first that we work only with choices of representatives  $\xi_{ij}$  which all share the same first coordinate  $x_i$ . If the partition of the interval  $[a, b]$  is fixed, and only the partition of  $[c, d]$  is refined, the values of the inner sum of the expression approaches the value of the integral

$$S_i = \int_{\varphi(x_i)}^{\eta(x_i)} f(x_i, y) \, dy,$$

which exists since the function  $f(x_i, y)$  is continuous. In this way, a function is obtained which is continuous in the free parameter  $x_i$ , see 8.2.1. Therefore, further refinement of the partition of the interval  $[a, b]$  leads, in the limit, to the desired formula

$$\sum_i S_i \Delta x_i \rightarrow S = \int_a^b \left( \int_{\psi(x)}^{\eta(x)} f(x, y) \, dy \right) dx.$$

It remains to deal with the case of general choices of representatives of general divisions  $\Xi$ . Since  $M$  is clearly compact (bounded and closed by definition), and  $f$  is a continuous function on  $M$ , it is uniformly continuous there. Therefore, if a small real number  $\varepsilon > 0$  is selected beforehand, there is always a bound  $\delta > 0$  for the norm of the partitions, so that the values of the function  $f$  for the general choices  $\xi_{ij}$  differ by no more than  $\varepsilon$  from the choices used above. Thus, the limit processes results in the same value for general Riemann sums  $S_{\Xi, \xi}$  as seen above.

Now, the general case can be proved easily by induction.

In the case of  $n = 1$ , the result is trivial. The presented reasoning can easily be transformed for a general induction step, writing  $(x_2, \dots, x_n)$  instead of  $y$ , having  $x_1$  instead of  $x$ , and perceiving the particular little cubes of the divisions as  $(n - 1)$ -dimensional cubes Cartesian-multiplied by the last interval. In the last-but-one step of the proof, the induction hypothesis is used, rather than the simple one-dimensional integration. The final argument about uniform continuity remains the same. It is advised to write this proof in detail as an exercise. □

**8.2.5. Fubini theorem.** The latter theorem has a particularly simple shape in the case of a multidimensional interval  $M$ . Then all the functions in bounds for integration are just the constant bounds from the definition of  $M$ . But this means that the integration process can be carried out coordinate by coordinate in

□

**8.I.15.** Find the volume of the solid which lies inside the cylinder  $y^2 + z^2 = 4$  and the half-space  $x \geq 0$  and is bounded by the surface  $y^2 + z^2 + 2x = 16$ .

**Solution.** In cylindrical coordinates,

$$V = \int_0^{2\pi} \int_0^2 \int_0^{8-\frac{r^2}{2}} r \, dx \, dr \, d\varphi = 28\pi. \quad \square$$

**8.I.16. The centroid of a solid.** The coordinates  $(x_t, y_t, z_t)$  of the centroid of a (homogeneous) solid  $T$  with volume  $V$  in  $\mathbb{R}^3$  are given by the following integrals:

$$\begin{aligned} x_t &= \iiint_T x \, dx \, dy \, dz, \\ y_t &= \iiint_T y \, dx \, dy \, dz, \\ z_t &= \iiint_T z \, dx \, dy \, dz. \end{aligned}$$

The centroid of a figure in  $\mathbb{R}^2$  or other dimensions can be computed analogously.

**8.I.17.** Find the centroid of the part of the ellipse  $3x^2 + 2y^2 = 1$  which lies in the first quadrant of the plane  $\mathbb{R}^2$ .

**Solution.** First, let us calculate the volume of the given ellipse. The transformation  $x = \frac{1}{\sqrt{3}}x', y = \frac{1}{\sqrt{2}}y'$  with Jacobian  $\frac{1}{\sqrt{6}}$  leads to

$$\begin{aligned} S &= \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} dy \, dx = \frac{1}{\sqrt{6}} \int_0^1 \int_0^{\sqrt{1-x^2}} dy' \, dx' \\ &= \frac{\pi}{4\sqrt{6}}. \end{aligned}$$

The other integrals we need can be computed directly in Cartesian coordinates  $x$  and  $y$ :

$$\begin{aligned} T_x &= \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} x \, dy \, dx = \int_0^{\frac{1}{\sqrt{3}}} x \sqrt{\frac{1-3x^2}{2}} \, dx = \\ &= \frac{1}{2} \int_0^{\frac{1}{\sqrt{3}}} \sqrt{\frac{1-3t}{2}} \, dt = \frac{\sqrt{2}}{18}, \end{aligned}$$

$$\begin{aligned} T_y &= \int_0^{\frac{1}{\sqrt{3}}} \int_0^{\sqrt{\frac{1-3x^2}{2}}} y \, dy \, dx = \frac{1}{2} \int_0^{\frac{1}{\sqrt{3}}} \frac{1-3x^2}{2} \, dx = \\ &= \frac{1}{4} \int_0^{\frac{1}{\sqrt{3}}} (1-3x^2) \, dx = \frac{\sqrt{3}}{18}. \end{aligned}$$

Therefore, the coordinates of the centroid are  $[\frac{4\sqrt{3}}{9\pi}, \frac{2\sqrt{2}}{\pi}]$ .  $\square$

**8.I.18.** Find the volume and the centroid of a homogeneous cone of height  $h$  and circular base with radius  $r$ .

**Solution.** Positioning the cone so that the vertex is at the origin and points downwards, we have in cylindrical coordinates that

$$V = 4 \int_0^{\pi/2} \int_0^r \int_{\frac{h}{r}\rho}^h \rho \, dz \, d\rho \, d\varphi = \frac{1}{3}\pi hr^2.$$

any order. We have exploited this behavior already in Chapter 6, cf. 6.3.12. In this way is proved the important corollary:<sup>5</sup>

FUBINI THEOREM

**Theorem.** Every continuous function  $f(x_1, \dots, x_n)$  on a multidimensional interval  $M = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n]$  is Riemann integrable on  $M$ , and its integral

$$\begin{aligned} \int_M f(x_1, \dots, x_n) \, dx_1 \dots dx_n \\ = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n) \, dx_1 \dots dx_n \end{aligned}$$

is independent of the order in which the multiple integration is performed.

The possibility of changing the order of integration in multiple integrals is extremely useful. We have already taken advantage of this result, namely when studying the relation of Fourier transforms and convolutions, see paragraph 7.1.9.

**8.2.6. Unbounded regions and functions.** There is no simple concept of an improper integral for unbounded multivariate functions. The following example of multiple integration of an unbounded function is illustrative in this direction:

$$\begin{aligned} \int_0^1 \left( \int_0^1 \frac{x-y}{(x+y)^3} \, dy \right) dx &= \frac{1}{2} \\ \int_0^1 \left( \int_0^1 \frac{x-y}{(x+y)^3} \, dx \right) dy &= -\frac{1}{2}. \end{aligned}$$

The reason can be understood by looking at the properties of non-absolutely converging series. There, rearranging the summands can lead to an arbitrary result.

The situation is better if the Riemann integral of a bounded non-negative function  $f(x) \geq 0$  with non-compact support over the whole  $\mathbb{R}^n$  is calculated. Of course some extra information is needed on the decay of the function  $f$  for large arguments. For example, if  $f$  is Riemann integrable over each  $n$ -dimensional interval  $I$  and there is a universal bound

$$\int_I |f(x)| \, dx \leq C$$

with a constant  $C$  independent of the choice of the  $n$ -dimensional interval  $I \subset \mathbb{R}^n$ , then we may define

$$\int_{\mathbb{R}^n} f(x) \, dx = \lim_{r \rightarrow \infty} \int_{I_r} f(x) \, dx,$$

where  $I_r = \{(x_1, \dots, x_n); |x_j| < r, j = 1, \dots, n\}$ . The resulting limit, if it exists, is bounded by the same constant  $C$ .

<sup>5</sup>Guido Fubini (1907-1943) was an important Italian mathematician active also in applied areas of mathematics. Simple derivation of Fubini theorem builds upon the simple properties of Riemann integration and the continuity of the integrated function. Fubini, in fact, proved this result in a much more general context of integration, while the theorem just introduced was used by mathematicians like Cauchy at least a century before Fubini.

Apparently, the centroid lies on the  $z$ -axis. For the  $z$ -coordinate, we get

$$z = \frac{1}{V} \int_{\text{cone}} z dV = \frac{1}{V} \int_0^{\pi/2} \int_0^r \int_{\frac{h}{r}\rho}^h z \rho dz d\rho d\varphi = \frac{3}{4}h.$$

Thus, the centroid lies  $\frac{1}{4}h$  over the center of the cone's base.  $\square$

**8.I.19.** Find the centroid of the solid which is bounded by the paraboloid  $2x^2 + 2y^2 = z$ , the cylinder  $(x+1)^2 + y^2 = 0$ , and the plane  $z = 0$ .

**Solution.** First, we will compute the volume of the given solid. Again, we use the cylindric coordinates  $(x = r \cdot \cos \varphi, y = r \cdot \sin \varphi, z = z)$ , where the equation of the paraboloid is  $z = 2r^2$  and the equation of the cylinder reads  $r = -2 \cos(\varphi)$ . Moreover, taking into account the fact that the plane  $x = 0$  is tangent to the given cylinder, we can easily determine the bounds of the integral that corresponds to the volume of the examined solid:

$$\begin{aligned} V &= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2 \cos \varphi} \int_0^{2r^2} r dz dr d\varphi \\ &= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2 \cos \varphi} 2r^3 dr d\varphi \\ &= \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} 8 \cos^4 \varphi d\varphi = 3\pi, \end{aligned}$$

where the last integral can be computed using the method of recurrence from 6.2.6.

Now, let us find the centroid. Since the solid is symmetric with respect to the plane  $y = 0$ , the  $y$ -coordinate of the centroid must be zero. Then, the remaining coordinates  $x_T$  and  $z_T$  of the centroid can be computed by the following integrals:

$$\begin{aligned} x_T &= \frac{1}{V} \int \int \int_{\text{B}} x dx dy dz \\ &= \frac{1}{V} \int_0^{2r^2} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2 \cos \varphi} r^2 \cos \varphi dz dr d\varphi \\ &= \frac{1}{V} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2 \cos \varphi} 2r^4 \cos \varphi dr d\varphi \\ &= \frac{1}{V} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} -\frac{64}{5} \cos^6 \varphi d\varphi = -\frac{4}{3}, \end{aligned}$$

where the last integral was computed by 6.2.6 again.

Analogously for the  $z$ -coordinate of the centroid:

$$z_T = \frac{1}{V} \int_0^{2r^2} \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} \int_0^{-2 \cos \varphi} zr \cos \varphi dz dr d\varphi = \frac{20}{9}.$$

The coordinates of the centroid are thus  $[-\frac{4}{3}, 0, \frac{20}{9}]$ .  $\square$

In this case, the Fubini theorem is true in the form

$$\int_{\mathbb{R}^n} f(x) dx = \int_{-\infty}^{\infty} \dots \left( \int_{-\infty}^{\infty} f(x) dx_1 \right) \dots dx_n.$$

**8.2.7. Further remarks on integration.** The Riemann

integral of multivariate functions behaves even worse than in the case of functions of one variable in the sixth chapter. Therefore, more sophisticated approaches to integrations have been developed. They are mainly based on the concept of the measure of a set. We consider this problem briefly now.



As we shall see in 8.2.10, the Riemann integration of the indicator functions  $\chi_M$  of sets  $M \subset \mathbb{R}^n$  leads to a finitely additive measure. In probability theory in chapter 10, even elementary problems require a concept of measure which is additive over countable systems of disjoint sets. Having such a measure, measurable functions  $f$  can be defined by the condition that their preimages of bounded intervals,  $f^{-1}([a, b])$ , are measurable sets and the integral is built by approximation via such "horizontal strips", see the illustration. This is the starting point of Lebesgue integration.

add a regular diagram on Lebesgue integration

We omit further details here, but note the Riesz representation theorem<sup>6</sup> saying that for each linear functional  $I$  (i.e. a linear mapping valued in  $\mathbb{R}$ ) on continuous functions with compact support on a metric space  $X$ , there is the unique measure (with certain regularity properties) such that the integral associated to this measure extends  $I$ . In the case of the Riemann integral  $I$  on functions on  $\mathbb{R}^n$  this provides the Lebesgue measure and the Lebesgue integral.

Another point of view is the completion procedure for metric spaces. Consider the vector space  $X = \mathcal{S}_c(\mathbb{R}^n)$  of all continuous functions with compact support. It can be equipped with the  $L_p$  norms, similar to the univariate case from the seventh chapter, i.e.

$$\|f\|_p = \left( \int_{\mathbb{R}^n} |f(x_1, \dots, x_n)|^p dx_1 \dots dx_n \right)^{1/p}$$

for any  $1 \leq p < \infty$ . Since the Riemann integral is defined again in terms of partitions and the representative values, the properties of the norm can be verified in the same way as for univariate functions, using Hölder's and Minkowski's inequalities.

There are the metrics  $\|\cdot\|_p$  on  $X$ . The general theory provides its completion  $\tilde{X}$ , unique up to isometry, and it can be shown that it is again a space of functions. The Lebesgue integral mentioned above defines exactly these norms. Hence the spaces of functions with Lebesgue integrable powers  $|f|^p$  are obtained.

<sup>6</sup>Frigyes Riesz (1880-1956) was a famous Hungarian mathematician active in particular in functional analysis. He introduced this theorem in the special case of  $X$  being an interval in  $\mathbb{R}^n$  in 1909

**8.I.20.** Find the centroid of the homogeneous solid in  $\mathbb{R}^3$  which lies between the planes  $z = 0$  and  $z = 2$ , bounded by the cones  $x^2 + y^2 = z^2$  and  $x^2 + y^2 = 2z^2$ .

**Solution.** The problem can be solved in the same way as the previous ones. It would be advantageous to work in cylindrical coordinates.

However, we can notice that the solid in question is an “annular cone”: it is formed by cutting out a cone  $K_1$  with base radius 4 of a cone  $K_2$  with base radius 8, of common height 2.

The centroid of the examined solid can be determined by the “rule of lever”: the centroid of a system of two solids is the weighted arithmetic mean of of the particular solids’ centroids, weighed by the masses of the solids. We found out in exercise 8.I.18 that the centroid of a homogeneous cone is situated at quarter its height. Therefore, the centroids of both cones lie at the same point, and this points thus must be the centroid of the examined solid as well. Hence, the coordinates of the wanted centroid are  $[0, 0, \frac{3}{2}]$ .  $\square$

**8.I.21.** Find the volume of the solid in  $\mathbb{R}^3$  which is bounded by the cone part  $x^2 + y^2 = (z - 2)^2$  and the paraboloid  $x^2 + y^2 = 4 - z$ .

**Solution.** We build the corresponding integral in cylindrical coordinates, which evaluates as follows:

$$V = \int_0^{2\pi} \int_0^1 \int_{r+2}^{4-r^2} r \, dz \, dr \, d\varphi = \frac{5}{6}\pi.$$

**8.I.22.** Find the volume of the solid in  $\mathbb{R}^3$  which lies under the cone  $x^2 + y^2 = (z - 2)^2$ ,  $z \leq 2$  and over the paraboloid  $x^2 + y^2 = z$ .

**Solution.**

$$V = \int_0^{2\pi} \int_0^1 \int_{r^2}^{2-r} r \, dz \, dr \, d\varphi = \frac{5}{6}\pi.$$

**8.2.8. Change of coordinates.** When calculating integrals of univariate functions, the “substitution method” is used as one of the powerful tools, cf. 6.2.5. The method works similarly in the case of functions of more variables, when understanding its geometric meaning.



Recall and reinterpret the univariate case. There, the integrated expression  $f(x) \, dx$  infinitesimally describes the two-dimensional area of the rectangle whose sides are the (linearized) increment  $\Delta x$  of the variable  $x$ , i.e. the one-dimensional rectangle, and the value  $f(x)$ . If the variable  $x$  is transformed by the relation  $x = u(t)$ , then the linearized increment can be expressed with the help of the differential as

$$dx = \frac{du}{dt} dt,$$

and so the corresponding contribution for the integral is given by

$$f(u(t)) \frac{du}{dt} dt.$$

Here one either supposes that the sign of the derivative  $u'(t)$  is positive, or one interchanges the bounds of the integral, so that the sign does not effect the result.

Intuitively, the procedure for  $n$  variables should be similar. It is only necessary to recall the formula for (change of) the volume of parallelepipeds. The Riemann integrals are approximated by Riemann sums, which are based on the  $n$ -dimensional volume (area) of small multidimensional intervals  $\Delta x_{i_1 \dots i_n}$  in the variables, multiplied by the values of the function at the representative points  $\xi_{i_1 \dots i_n}$ . If the coordinates are transformed by means of a mapping  $x = G(y)$ , not only the function values  $f(G(\xi_{i_1 \dots i_n}))$  are obtained at the representative points  $\tilde{\xi}_{i_1 \dots i_n} = G^{-1}(\xi_{i_1 \dots i_n})$  in the new coordinate expression, but also the change of the volume of the corresponding small multidimensional intervals needs care.

Once again, this is the case of a linear approximation of a change, which is well known — the best linear approximation of  $G(y)$  is its derivative  $D^1 G(y)$ , which is given by the Jacobi matrix of  $G$ , see 8.1.19. The change of the volume is then given (in absolute value) by the determinant of this matrix (see a discussion of this topic in chapter 4 devoted to analytic geometry and linear algebra, especially 4.1.19).

Summarizing, the formulation of the next theorem should not be surprising and its proof consists in formalization of the latter ideas. However, this needs some effort and so the proof is split into several steps.



Note that the considered solid is symmetric with the solid from the previous exercise 8.I.21 (the center of the symmetry is the point  $[0, 0, 2]$ ). Therefore, it must have the same volume.  $\square$

8.I.23. Find the centroid of the surface bounded by the parabola  $y = 4 - x^2$  and the line  $y = 0$ .  $\circ$

8.I.24. Find the centroid of the circular sector corresponding to the angle of  $60^\circ$  that was cut out of a disc with radius 1.  $\circ$

8.I.25. Find the centroid of the semidisc  $x^2 + y^2 = 1, y \geq 0$ .  $\circ$

8.I.26. Find the centroid of the circular sector corresponding to the angle of  $120^\circ$  that was cut out of a disc with radius 1.  $\circ$

8.I.27. Find the volume of the solid in  $\mathbb{R}^3$  which is given by the inequalities  $z \geq 0, z - x \leq 0$ , and  $(x - 1)^2 + y^2 \leq 1$ .  $\circ$

8.I.28. Find the volume of the solid in  $\mathbb{R}^3$  which is given by the inequalities  $z \geq 0, z - y \leq 0$ .  $\circ$

8.I.29. Find the volume of the solid bounded by the surface

$$3x^2 + 2y^2 + 3z^2 + 2xy - 2yz - 4xz = 1.$$

8.I.30. Find the volume of the part of  $\mathbb{R}^3$  lying inside the ellipsoid  $2x^2 + y^2 + z^2 = 6$  and in the half-space  $x \geq 1$ .  $\circ$

8.I.31. **The area of the graph of a real-valued function  $f(x, y)$  in variables  $x$  and  $y$ .** The area of the graph of a function of two variables over an area  $S$  in the plane  $xy$  is given by the integral

$$P = \int_S \sqrt{1 + f_x^2 + f_y^2} \, dx \, dy.$$

Considering the cone  $x^2 + y^2 = z^2$ , find the area of the part of its lateral surface which lies above the plane  $z = 0$  and inside the cylinder  $x^2 + y^2 = y$ .

**Solution.** The wanted area can be calculated as the area of the graph of the function  $z = \sqrt{x^2 + y^2}$  over the disc  $K: x^2 - (y - \frac{1}{2})^2$ . We can easily see that

$$f_x = \frac{x}{x^2 + y^2}, \quad f_y = \frac{y}{x^2 + y^2},$$

so the area is expressed by the integral

$$\begin{aligned} \iint_K \sqrt{1 + f_x^2 + f_y^2} \, dx \, dy &= \iint_K \sqrt{2} \, dx \, dy = \\ &= \sqrt{2} \int_0^\pi \int_0^{\sin \pi} r \, dr \, d\varphi = \frac{\sqrt{2}}{2} \int_0^\pi \sin^2 \varphi \\ &= \frac{\sqrt{2}\pi}{4}. \end{aligned}$$

**Theorem.** Let  $G(t) : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a continuously differentiable and invertible mapping, and write

$$t = (t_1, \dots, t_n), \quad x = (x_1, \dots, x_n) = G(t_1, \dots, t_n).$$

Further let  $M = G(N)$  be a Riemann-measurable sets, and  $f : M \rightarrow \mathbb{R}$  a continuous function. Then,  $N$  is also Riemann-measurable and

$$\int_M f(x) \, dx_1 \dots dx_n = \int_N f(G(t)) |\det(D^1 G(t))| \, dt_1 \dots dt_n.$$

**8.2.9. The invariance of the integral.** The first thing to be verified is the coincidence of two definitions of volume of parallelepipeds (taken for granted in the above intuitive explanation of the latter theorem). Volumes and similar concepts were dealt with in chapter 4 and a crucial property was the invariance of the concepts with respect to the choice of Euclidean frames of  $\mathbb{R}^n$ , cf. 4.1.19 on page 334, which followed directly from the expression of the volumes in terms of determinants. It is needed to show that the same result holds in terms of the Riemann integration as defined above. It turns out that it is easier to deal with invariance with respect to general invertible linear mappings  $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ .

**Proposition.** Let  $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be an invertible linear mapping and  $I \subset \mathbb{R}^n$  a multidimensional interval. Consider a function  $f$ , such that  $f \circ \Psi$  is integrable on  $I$ . Then  $M = \Psi(I)$  is Riemann-measurable,  $f$  is Riemann integrable on  $M$  and

$$\begin{aligned} \int_M f(x_1, \dots, x_n) \, dx_1 \dots dx_n &= \\ &= |\det \Psi| \int_I (f \circ \Psi)(y_1, \dots, y_n) \, dy_1 \dots dy_n. \end{aligned}$$

**PROOF.** Each linear mapping is a composition of the elementary transformations of three types (see the discussion in chapter 2, in particular paragraphs 2.1.7 and 2.1.9).

The first one is a multiplication of one of the coordinates with a constant:  $\Psi(y_1, \dots, y_n) = (y_1, \dots, \alpha y_i, \dots, y_n)$ . In this case  $|\det \Psi| = |\alpha|$ . The second one consists of an exchange of two coordinates, i.e. for given  $1 \leq i < j \leq n$ ,  $\Psi(y_1, \dots, y_n) = (y_1, \dots, y_j, \dots, y_i, \dots, y_n)$ . The determinant of  $\Psi$  is  $-1$  in this case. The third type of transformations is of the form  $\Psi(y_1, \dots, y_n) = (y_1, \dots, y_i + y_j, \dots, y_j, \dots, y_n)$ , with determinant one. Without loss of generality,  $i = 1$  can be chosen in the first case and  $i = 1, j = 2$ , in the second case. Since the determinant of the composition of the mappings (i.e. the determinant of the product of the matrices) is the product of the individual determinants, it is enough to prove the proposition for all three special types of  $\Psi$ .

Express the right hand integrals for these three types of  $\Psi$  by means of the multiple integrals and Fubini theorem. Write

**8.I.32.** Find the area of the parabola  $z = x^2 + y^2$  over the disc  $x^2 + y^2 \leq 4$ .  $\square$

**8.I.33.** Find the area of the part of the plane  $x + 2y + z = 10$  that lies over the figure given by  $(x - 1)^2 + y^2 \leq 1$  and  $y \geq x$ .  $\square$

In the following exercise, we will also apply our knowledge of the theory of Fourier transforms from the previous chapter.

**8.I.34. Fourier transform and diffraction.** Light intensity is a physical quantity which expresses the transmission of energy by waves. The intensity of a general light wave is defined as the time-averaged magnitude of the Poynting vector, which is the vector product of mutually orthogonal vectors of electric and magnetic fields. A monochromatic plane wave spreading in the direction of the  $y$ -axis satisfies

$$I = c\varepsilon_0 \frac{1}{\tau} \int_0^\tau E_y^2 dt,$$

where  $c$  is the speed of light and  $\varepsilon_0$  is the vacuum permittivity. The monochromatic wave is described by the harmonic function  $E_y = \psi(x, t) = A \cos(\omega t - kx)$ . The number  $A$  is the maximal amplitude of the wave,  $\omega$  is the angular frequency, and for any fixed  $t$ , the so-called wave length  $\lambda$  is the prime period. The number  $k$  then represents the speed  $k = \frac{2\pi}{\lambda}$  at which the wave propagates. We have

$$\begin{aligned} I &= c\varepsilon_0 \frac{1}{\tau} \int_0^\tau E_y^2 dt = c\varepsilon_0 \frac{1}{\tau} \int_0^\tau A^2 \cos^2(\omega t - kx) dt \\ &= c\varepsilon_0 A^2 \frac{1}{\tau} \int_0^\tau \frac{1 + \cos(2(\omega t - kx))}{2} dt \\ &= \frac{1}{2} c\varepsilon_0 A^2 \frac{1}{\tau} \left[ t + \frac{\sin(2(\omega t - kx))}{2\omega} \right]_0^\tau \\ &= \frac{1}{2} c\varepsilon_0 A^2 \frac{1}{\tau} \left( \tau + \frac{\sin(2(\omega\tau - kx)) - \sin(2(-kx))}{2\omega} \right) \\ &= \frac{1}{2} c\varepsilon_0 A^2 \left( 1 + \frac{\sin(2(\omega\tau - kx)) - \sin(2(-kx))}{2\omega\tau} \right) \\ &\doteq \frac{1}{2} c\varepsilon_0 A^2 \end{aligned}$$

The second term in the parentheses can be neglected since it is always less than  $\frac{2}{2\omega\tau} = \frac{T}{2\pi\tau} < 10^{-6}$  for real detectors of light, so it is much inferior to 1. The light intensity is directly proportional to the squared amplitude.

A diffraction is such a deviation from straight-line propagation of light which cannot be explained as the result of a refraction or reflection (or the change of the ray's direction in a medium with continuously varying refractive index). The diffraction can be observed when a lightbeam propagates through a bounded space. The diffraction phenomena are strongest and easiest to see if the light goes through openings or obstacles whose size is roughly the wavelength of the light. In the case of the Fraunhofer diffraction, with which we will deal in the following example, a monochromatic plane wave goes through a very thin rectangular opening and projects on

$I = [a_1, b_1] \times \dots \times [a_n, b_n]$  and  $x = \Psi(y)$  for the transformation. In the first case (notice we can deal with the first variable and  $\alpha > 0$  without loss of generality),

$$\begin{aligned} |\det \Psi| \int_I f(\alpha y_1, y_2, \dots, y_n) dy_1 \dots dy_n &= \\ &= \alpha \int_{a_n}^{b_n} \dots \left( \int_{a_1}^{b_1} f(\alpha y_1, y_2, \dots, y_n) dy_1 \right) \dots dy_n \\ &= \alpha \alpha^{-1} \int_{a_n}^{b_n} \dots \left( \int_{\alpha a_1}^{\alpha b_1} f(x_1, x_2, \dots, x_n) dx_1 \right) \dots dx_n \\ &= \int_{\Psi(I)} f(x_1, x_2, \dots, x_n) dx_1 \dots dx_n. \end{aligned}$$

The second case is even easier, since the order of integration does not matter due to the Fubini theorem. The third case is similar to the first one:

$$\begin{aligned} |\det \Psi| \int_I f(y_1 + y_2, y_2, \dots, y_n) dy_1 \dots dy_n &= \\ &= \int_{a_n}^{b_n} \dots \left( \int_{a_1}^{b_1} f(y_1 + y_2, y_2, \dots, y_n) dy_1 \right) \dots dy_n \\ &= \int_{a_n}^{b_n} \dots \left( \int_{a_1 + x_2}^{b_1 + x_2} f(x_1, x_2, \dots, x_n) dx_1 \right) \dots dx_n \\ &= \int_{\Psi(I)} f(x_1, x_2, \dots, x_n) dx_1 \dots dx_n. \end{aligned}$$

The reader should check the details that the last multiple integral describes the image  $\Psi(I)$ .  $\square$

As a direct corollary of the proposition, the Riemann integral is invariant with respect to the Euclidean affine mappings. That is, the integral cannot depend on the choice of the orthogonal frame in the Euclidean  $\mathbb{R}^n$ .

**8.2.10. Riemann-measurable sets.** It is necessary to understand how to recognize Riemann-measurable domains  $M$ .

When defining the Riemann integral, a strict analogy of the lower and upper Riemann integrals for univariate functions can be considered. This means taking infima or suprema of the integrated function over the corresponding multidimensional intervals instead of the function values at the representatives in the Riemann sums. For bounded functions, there are well-defined values of the upper and lower integrals found in this way. If this is done for the indicator function  $\chi_M$  of a fixed set  $M$ , the *inner and outer Riemann measure* of the set  $M$  is obtained. Evidently, the inner measure is the supremum of the areas given by the (finite) sums of the volumes of all multi-dimensional intervals from the partitions which are inside  $M$ , and on the other hand, the outer measure is the infimum of the (finite) sums of the volumes of intervals covering  $M$ . It follows directly from the definition that a set  $M$  is *Riemann-measurable* if and only if its inner and outer measures are equal.

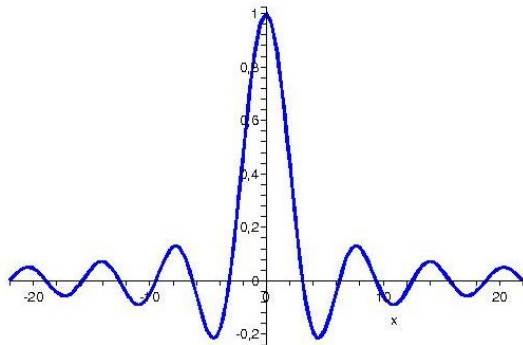
The sets whose outer measure is zero are, of course, Riemann-measurable. They are called *measure zero sets* or null sets. The finite additivity of the Riemann integral makes

a distant surface. For instance, we can highlight a spot on the wall with a laser pointer. The image we get is the Fourier transform of the function describing the permeability of the shade - opening.

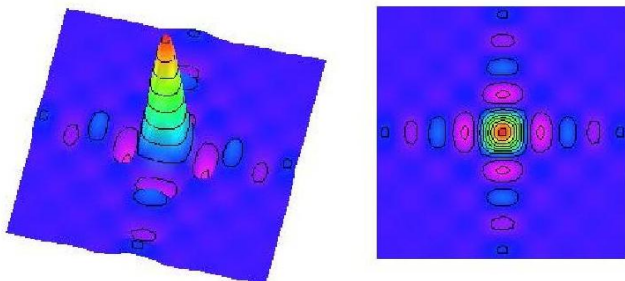
Let us choose the plane of the diffraction shade as the coordinate plane  $z = 0$ . Let a plane wave  $A \exp(ikz)$  (independent of the point  $(x, y)$  of landing on the shade) hit this plane perpendicularly. Let  $s(x, y)$  denote the function of the permeability of the shade, then the resulting waves falling onto the projection surface at a point  $(\xi, \eta)$  can be described as the integral sum of the waves (Huygens-Fresnel principle) which have gone through the shade and propagate through the medium from all points  $(x, y, 0)$  (as a spherical wave) into the point  $(\xi, \eta, z)$ :

$$\begin{aligned} \psi(\xi, \eta) &= A \iint_{\mathbb{R}^2} s(x, y) e^{-ik(\xi x + \eta y)} dx dy \\ \psi(\xi, \eta) &= A \int_{-p/2}^{p/2} \int_{-q/2}^{q/2} e^{-ik(\xi x + \eta y)} dy dx \\ \psi(\xi, \eta) &= A \int_{-p/2}^{p/2} e^{-ik\xi x} dx \int_{-q/2}^{q/2} e^{-ik\eta y} dy \\ &= A \left[ \frac{e^{-ik\xi x}}{-ik\xi} \right]_{-p/2}^{p/2} \left[ \frac{e^{-ik\eta y}}{-ik\eta} \right]_{-q/2}^{q/2} \\ &= A \frac{2 \sin(k \xi p/2)}{k\xi} \frac{2 \sin(k \eta q/2)}{k\eta} \\ &= Apq \frac{\sin(k \xi p/2)}{k\xi p/2} \frac{\sin(k \eta q/2)}{k\eta q/2} \end{aligned}$$

The graph of the function  $f(x) = \frac{\sin x}{x}$  looks as follows:



The graph of the function  $\psi(\xi, \eta) = \frac{\sin \xi}{\xi} \frac{\sin \eta}{\eta}$  then does:



the measure finitely additive. Hence, a disjoint union of finitely many measurable sets is again a measurable set, and its measure is given by the sum of the measures of the individual sets in the union.

Consider the measurability of any given set  $M \subset I \subset \mathbb{R}^n$  inside a sufficiently large multidimensional interval  $I$ . Consider the boundary  $\partial M$ , i.e. the set of all boundary points of  $M$ . For any partition  $\Xi$  of  $I$  from the definition of the Riemann integral of  $\chi_M$ , each of the intervals  $I_{i_1 \dots i_n}$  with non-trivial intersection with  $\partial M$  contributes to the upper integral but might not contribute to the lower integral. On the contrary, for every point in the interior  $M_o \subset M$  its interval  $I_{i_1 \dots i_n}$  contributes to both the same way as soon as the norm of the partition is small enough. This observation leads to the first part of the following claim:

**Proposition.** *A bounded set  $M \subset \mathbb{R}^n$  is Riemann-measurable if and only if its boundary is of Riemann-measure zero.*

*If  $M$  is a Riemann-measurable set and  $G : M \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a continuously differentiable and invertible mapping, then  $G(M)$  is again Riemann-measurable.*

**PROOF.** The first claim is already verified. Since both  $G$  and  $G^{-1}$  are continuous,  $G$  maps internal points of  $M$  to internal points of  $G(M)$ . To finish the proof, it must be verified that  $G$  maps the boundary  $\partial M$ , which is a set of measure zero, again to a set of measure zero.

Since every Riemann integrable set  $M$  is bounded, its closure  $\bar{M}$  must be compact. It follows that  $G$  and all partial derivatives of its components are uniformly continuous on  $\bar{M}$ , and in particular on the boundary  $\partial M$ .

Next, consider a partition  $\Xi$  of an interval  $I$  containing  $\partial M$  and a fixed tiny interval  $J$  in the partition including a point  $t \in \partial M$ . Write  $R = G(t) + D^1G(t)(J - t)$ , which is to be understood as follows:  $J$  is first shifted to the origin by translation, then the derivative of  $G$  is applied obtaining a parallelepiped. This is shifted back to be around  $G(t)$ . By the uniform continuity of  $G$  and  $D^1G$ , for each  $\varepsilon > 0$  there is a bound  $\delta$  for the norm of a partition for which

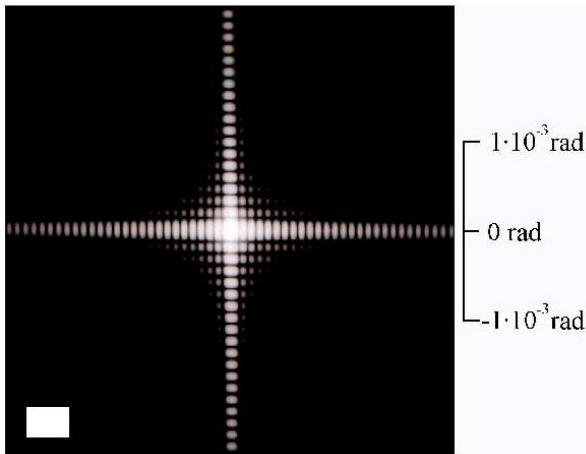
$$G(J) \subset G(t) + (1 + \varepsilon)D^1G(t)(J - t)$$

can be guaranteed. The entire image of  $J$  lies inside a slightly enlarged linear image of  $J$  by the derivative. Now, the outer measure  $\alpha$  of the image  $G(J)$  satisfies:

$$\alpha \leq (1 + \varepsilon)^n \text{vol}_n R = (1 + \varepsilon)^n |\det G(t)| \text{vol}_n J.$$

If  $\mu$  is the upper Riemann sum for the measure of  $\partial M$  corresponding to the chosen partition, the outer measure of  $G(\partial M)$  must be bounded by  $(1 + \varepsilon)^n \max_{t \in \partial M} |\det G(t)| \mu$ . Finally, we exploit the assumption that  $\partial M$  has got measure zero and thus, for the same  $\varepsilon$  we may further decrease the bound on the norm of the partition so that  $\mu \leq \varepsilon$ , too. But then the outer measure is bounded by a constant multiple of  $(1 + \varepsilon)^n \varepsilon$ , with the universal constant  $\max_{t \in \partial M} |\det G(t)|$ . So the outer measure is zero, as required.  $\square$

And the diffraction we are describing:



Since  $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$ , the intensity at the middle of the image is directly proportional to  $I_0 = A^2 p^2 q^2$ . The Fourier transform can be easily scrutinized if we aim a laser pointer through a subtle opening between the thumb and the index finger; it will be the image of the function of its permeability. The image of the last picture can be seen if we create a good rectangular opening by, for instance, gluing together some stickers with sharp edges.

### J. First-order differential equations

**8.J.1.** Find all solutions of the differential equation

$$y' = \frac{\sqrt{1-y^2}}{\cos^2 x} (1 + \cos^2 x).$$

**Solution.** We are given an ordinary first-order differential equation in the form  $y' = f(x, y)$ , which is called an explicit form of the equation. Moreover, we can write it as  $y' = f_1(x) \cdot f_2(y)$  for continuous univariate functions  $f_1$  and  $f_2$  (on certain open intervals), i. e., it is a differential equation with separated variables.

First, we replace  $y'$  with  $dy/dx$  and rewrite the differential equation in the form

$$\frac{1}{\sqrt{1-y^2}} dy = \frac{1+\cos^2 x}{\cos^2 x} dx.$$

Since

$$\int \frac{1+\cos^2 x}{\cos^2 x} dx = \int \frac{1}{\cos^2 x} + 1 dx,$$

we can integrate using the basic formulae, thereby obtaining

$$(1) \quad \arcsin y = \operatorname{tg} x + x + C, \quad C \in \mathbb{R}.$$

However, we must keep in mind that the division by the expression  $\sqrt{1-y^2}$  is valid only if it is non-zero, i. e., only for  $y \neq \pm 1$ . Substituting the constant functions  $y \equiv 1$ ,  $y \equiv -1$  into the given differential equation, we can immediately see that they satisfy it. We have thus obtained two more solutions, which are called singular. We do not have to pay attention to the case  $\cos x = 0$  since this only loses points of the domains (but not any solutions).

Now, we will comment on several parts of the computation. The expression  $y' = dy/dx$  allows us to make many symbolic manipulations. For instance, we have

A slightly extended argumentation as in the proof above leads to understanding that the Riemann integrable functions are exactly those bounded functions with compact support whose set of discontinuity points has (Riemann) measure zero.

**8.2.11. Proof of Theorem 8.2.8.** A continuous function  $f$  and a differentiable change of coordinates is under consideration. So the inverse  $G^{-1}$  is continuously differentiable, and the image  $G^{-1}(M) = N$  is Riemann-measurable. Hence the integrals on both sides of the equality exist and it remains to prove that their values are equal.

Denote a composite continuous function by

$$g(t_1, \dots, t_n) = f(G(t_1, \dots, t_n)),$$

and choose a sufficiently large  $n$ -dimensional interval  $I$  containing  $N$  and its partition  $\Xi$ . The entire proof is nothing more than a more exact writing of the discussion presented before the formulation of the theorem.

Repeat the estimates on the volumes of images from the previous paragraph on Riemann measurability. It is already known that the images  $G(I_{i_1 \dots i_n})$  of the intervals from the partition are again Riemann-measurable sets. For each small part  $I_{i_1 \dots i_n}$  of the partition  $\Xi$ , the integral of  $f$  over  $J_{i_1 \dots i_n} = G(I_{i_1 \dots i_n})$  certainly exists, too.

Further, if the center  $t_{i_1 \dots i_n}$  of the interval  $I_{i_1 \dots i_n}$  is fixed, then the linear image of this interval

$$R_{i_1 \dots i_n} = G(t_{i_1 \dots i_n}) + D^1 G(t_{i_1 \dots i_n})(I_{i_1 \dots i_n} - t_{i_1 \dots i_n}),$$

is obtained. This is an  $n$ -dimensional parallelepiped (note that the interval is shifted to the origin, then mapped by the linear mapping given by the Jacobi matrix, and the result is then added to the image of the center).

If the partition is very fine, this parallelepiped differs only a little from the image  $J_{i_1 \dots i_n}$ . By the uniform continuity of the mapping  $G$ , there is, for an arbitrarily small  $\varepsilon > 0$ , a norm of the partition such that for all finer partitions

$$G(t_{i_1 \dots i_n}) + (1 + \varepsilon)D^1 G(t_{i_1 \dots i_n})(I_{i_1 \dots i_n}) \supset J_{i_1 \dots i_n}.$$

However, then the  $n$ -dimensional volumes also satisfy

$$\begin{aligned} \operatorname{vol}_n(J_{i_1 \dots i_n}) &\leq (1 + \varepsilon)^n \operatorname{vol}_n(R_{i_1 \dots i_n}) \\ &= (1 + \varepsilon)^n |\det G(t_{i_1 \dots i_n})| \operatorname{vol}_n(I_{i_1 \dots i_n}). \end{aligned}$$

Now, it is possible to estimate the entire integral:

$$\begin{aligned} \int_M f(x_1, \dots, x_n) dx_1 \dots dx_n &= \\ &= \sum_{i_1 \dots i_n} \int_{J_{i_1 \dots i_n}} f(x_1, \dots, x_n) dx_1 \dots dx_n \\ &\leq \sum_{i_1 \dots i_n} \left( \sup_{t \in I_{i_1 \dots i_n}} g(t) \right) \operatorname{vol}_n(J_{i_1 \dots i_n}) \\ &\leq (1 + \varepsilon)^n \sum_{i_1 \dots i_n} \left( \sup_{t \in I_{i_1 \dots i_n}} g(t) \right) |\det G(t_{i_1 \dots i_n})| \operatorname{vol}_n(I_{i_1 \dots i_n}). \end{aligned}$$

If  $\varepsilon$  approaches zero, then the norms of the partitions approach zero too, the left-hand value of the integral remains the same, while on the right-hand side, the Riemann integral

$$\frac{dz}{dy} \cdot \frac{dy}{dx} = \frac{dz}{dx}, \quad \frac{1}{\frac{dy}{dx}} = \frac{dx}{dy}.$$

The validity of these two formulae is actually guaranteed by the chain rule theorem and the theorem for differentiating an inverse function, respectively. It was just the facility of the manipulations that inspired G. W. Leibniz to introduce this notation, which has been in use up to now. Further, we should realize why we have not written the general solution (1) in the suggesting form

$$(2) \quad y = \sin(\operatorname{tg} x + x + C), \quad C \in \mathbb{R}.$$

As we will not mention the domains of differential equations (i. e., for which values of  $x$  the expressions are well-defined), we will not change them by “redundant” simplifications, either. It is apparent that the function  $y$  from (2) is defined for all  $x \in (0, \pi) \setminus \{\pi/2\}$ . However, for the values of  $x$  which are close to  $\pi/2$  (having fixed  $C$ ), there is no  $y$  satisfying (1). In general, the solutions of differential equations are curves which may not be expressible as graphs of elementary functions (on the whole intervals where we consider them). Therefore, we will not even try to do that.  $\square$

**8.J.2.** Find the general solution of the equation  $y' = (2 - y) \operatorname{tg} x$ .

**Solution.** Again, we are given a differential equation with separated variables.

We have

$$\begin{aligned} \frac{dy}{dx} &= (2 - y) \operatorname{tg} x, \\ -\frac{dy}{y - 2} &= \frac{\sin x}{\cos x} dx, \\ -\ln|y - 2| &= -\ln|\cos x| - \ln|C|, \quad C \neq 0. \end{aligned}$$

Here, the shift obtained from the integration has been expressed by  $\ln|C|$ , which is very advantageous (bearing in mind what we want to do next) especially in those cases when we obtain a logarithm on both sides of the equation. Further, we have

$$\begin{aligned} \ln|y - 2| &= \ln|C \cos x|, \quad C \neq 0, \\ |y - 2| &= |C \cos x|, \quad C \neq 0, \\ y - 2 &= C \cos x, \quad C \neq 0, \end{aligned}$$

where we should write  $\pm C$  (after removing the absolute value). However, since we consider all non-zero values of  $C$ , it makes no difference whether we write  $+C$  or  $-C$ . We should pay attention to the fact that we have made a division by the expression  $y - 2$ . Therefore, we must examine the case  $y \equiv 2$  separately. The derivative of a constant function is zero, so we have found another solution,  $y \equiv 2$ . However, this solution is not singular since it is contained in the general solution as the case  $C = 0$ . Thus, the correct result is

$$y = 2 + C \cos x, \quad C \in \mathbb{R}. \quad \square$$

of  $g(t)|\det G(t)|$  is obtained. Instead of the desired equality, the inequality:

$$\int_M f(x) dx_1 \dots dx_n \leq \int_N f(G(t)) |\det(D^1 G(t))| dt_1 \dots dt_n$$

is obtained.

The same reasoning can be repeated after interchanging  $G$  and  $G^{-1}$ , the integration domains  $M$  and  $N$ , and the functions  $f$  and  $g$ . The reverse inequality is immediately obtained:

$$\begin{aligned} &\int_N g(t) |\det(D^1 G(t))| dt_1 \dots dt_n \\ &\leq \int_M f(x) |\det(D^1 G(G^{-1}(x)))| \\ &\quad |\det(D^1 G^{-1}(x))| dx_1 \dots dx_n \\ &= \int_M f(x) dx_1 \dots dx_n. \end{aligned}$$

The proof is complete.

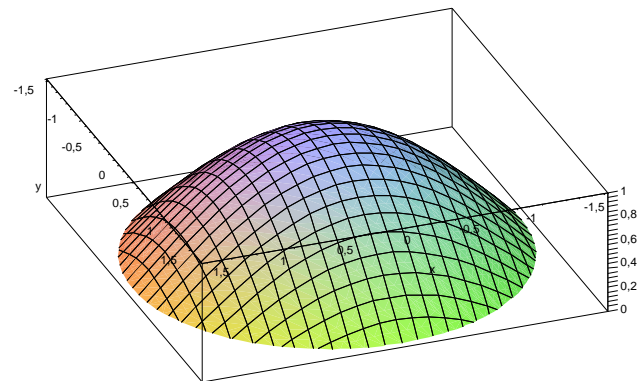
**8.2.12. An example in two dimensions.** The coordinate transformations are quite transparent for the integral of a continuous function  $f(x, y)$  of two variables. Consider the differentiable transformation  $G(s, t) = (x(s, t), y(s, t))$ . Denoting  $g(s, t) = f(x(s, t), y(s, t))$ ,



$$\int_{G(N)} f(x, y) dx dy = \int_N g(s, t) \left| \frac{\partial x}{\partial s} \frac{\partial y}{\partial t} - \frac{\partial x}{\partial t} \frac{\partial y}{\partial s} \right| ds dt$$

is obtained.

As a truly simple example, calculate the integral of the indicator function of a disc with radius  $R$  (i.e. its area) and the integral of the function  $f(t, \theta) = \cos t$  defined in polar coordinates inside a circle with radius  $\frac{1}{2}\pi$  (i.e. the volume hidden under such a “cap placed above the origin”, see the illustration).



First, determine the Jacobi matrix of the transformation  $x = r \cos \theta, y = r \sin \theta$

$$D^1 G = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}.$$

Hence, the determinant of this matrix is equal to

$$\det D^1 G(r, \theta) = r(\sin^2 \theta + \cos^2 \theta) = r.$$

**8.J.3.** Find the solution of the differential equation

$$(1 + e^x)yy' = e^x$$

which satisfies the initial condition  $y(0) = 1$ .

**Solution.** If the functions  $f : (a, b) \rightarrow \mathbb{R}$  and  $g : (c, d) \rightarrow \mathbb{R}$  are continuous and  $g(y) \neq 0$ ,  $y \in (c, d)$ , then the initial problem

$$y' = f(x)g(y), \quad y(x_0) = y_0$$

has a unique solution for any  $x_0 \in (a, b)$ ,  $y_0 \in (c, d)$ . This solution is determined implicitly as

$$\int_{y_0}^{y(x)} \frac{dt}{g(t)} = \int_{x_0}^x f(t) dt.$$

In practical problems, we first find all solutions of the equation and then select the one which satisfies the initial condition.

Let us compute:

$$(1 + e^x)y dy/dx = e^x,$$

$$y dy = \frac{e^x}{1 + e^x} dx,$$

$$\frac{y^2}{2} = \ln(1 + e^x) + \ln|C|, \quad C \neq 0,$$

$$\frac{y^2}{2} = \ln(C[1 + e^x]), \quad C > 0.$$

The substitution  $y = 1$ ,  $x = 0$  then gives

$$\frac{1}{2} = \ln(C \cdot 2), \quad \text{i. e. } C = \frac{\sqrt{e}}{2}.$$

We have thus found the solution

$$\frac{y^2}{2} = \ln\left(\frac{\sqrt{e}}{2}[1 + e^x]\right),$$

i. e.,

$$y = \sqrt{2 \ln\left(\frac{\sqrt{e}}{2}[1 + e^x]\right)}$$

on a neighborhood of the point  $[0, 1]$  where  $y > 0$ . □

**8.J.4.** Find the solution of the differential equation

$$y' = \frac{y^2+1}{x+1}$$

which satisfies  $y(0) = 1$ .

**Solution.** Similarly to the previous example, we get

$$\frac{dy}{y^2+1} = \frac{dx}{x+1},$$

$$\arctan y = \ln|x+1| + C, \quad C \in \mathbb{R}.$$

The initial condition (i. e., the substitution  $x = 0$  and  $y = 1$ ) gives

$$\arctan 1 = \ln|1| + C, \quad \text{i. e., } C = \frac{\pi}{4}.$$

Therefore, the solution of the given initial problem is the function

$$y(x) = \text{tg}\left(\ln|x+1| + \frac{\pi}{4}\right)$$

on a neighborhood of the point  $[0, 1]$ . □

Therefore, the calculation can be done directly for the disc  $S$  which is the image of the rectangle  $(r, \theta) \in [0, R] \times [0, 2\pi] = T$ . In this way the area of the disc is obtained:

$$\int_S dx dy = \int_0^{2\pi} \int_0^R r dr d\theta = \int_0^R 2\pi r dr = \pi R^2.$$

The integration of the function  $f$  is very similar, using multiple integration and integration by parts:

$$\int_S f(x, y) dx dy = \int_0^{2\pi} \int_0^{\pi/2} r \cos r dr d\theta = \pi^2 - 2\pi.$$

In many real life applications, a much more general approach to integration is needed which allows for the dealing with objects over curves, surfaces, and their higher dimensional analogues. For many simple cases, such tools can be built now with the help of parametrization of such  $k$ -dimensional surfaces and employ the letter theorem to show the independence of the result on such a parametrization. These topics are postponed to the beginning of the next chapter where a more general and geometric approach is discussed, see e.g. 9.1.9 and 9.1.12.

### 3. Differential equations

In this section, we return to (vector) functions of one variable, defined and examined in terms of their instantaneous changes.

**8.3.1. Linear and non-linear difference models.** The concept of derivative was introduced in order to work with instantaneous changes of the examined quantities. In the introductory chapter, difference equations based on similar concepts in relation to sequences of scalars were discussed. As a motivating introduction to equations containing derivatives of unknown functions, recall first the difference equations.



The simplest difference equations are formulated as  $y_{n+1} = F(y_n, n)$ , with a function  $F$  of two variables. For example, the model describing interests of deposits or loans (this included the Malthusian model of populations) was considered. The increment was proportional to the value,  $y_{n+1} = a y_n$ , see 1.2.2. Growths by 5% is represented by  $a = 1.05$ . Considering continuous modeling, the same request leads to an equation connecting the derivative  $y'(t)$  of a function with its value

(1) 
$$y'(t) = r y(t)$$

with the proportionality constant  $r$ . Here, the instantaneous growth by 5% corresponds to  $r = 0.05$ .

It is easy to guess the *solution* of the latter equation, i.e. a function  $y(t)$  which satisfies the equality identically,

$$y(t) = C e^{rt}$$

with an arbitrary constant  $C$ . This constant can be determined uniquely by choosing the *initial value*  $y_0 = y(t_0)$  at some point  $t_0$ . If a part of the increment in a model should be given as a constant independent of the value  $y$  or  $t$  (like bank charges

**8.J.5.** Solve

$$(1) \quad y' = \frac{x + y + 1}{2x + 2y - 1}.$$

**Solution.** Let a function  $f : (a, b) \times (c, d) \rightarrow \mathbb{R}$  have continuous second-order partial derivatives and  $f(x, y) \neq 0$ ,  $x \in (a, b)$ ,  $y \in (c, d)$ . Then, the differential equation  $y' = f(x, y)$  can be transformed to an equation with separated variables if and only if

$$\begin{vmatrix} f(x, y) & f'_y(x, y) \\ f'_x(x, y) & f''_{xy}(x, y) \end{vmatrix} = 0, \quad x \in (a, b), y \in (c, d).$$

With a bit of effort, it can be shown that a differential equation of the form  $y' = f(ax + by + c)$  can be transformed to an equation with separated variables, and this can be done by the substitution  $z = ax + by + c$ . Let us emphasize that the variable  $z$  replaces  $y$ .

We thus set  $z = x + y$ , which gives  $z' = 1 + y'$ . Substitution into (1) yields

$$\begin{aligned} z' - 1 &= \frac{z + 1}{2z - 1}, \\ \frac{dz}{dx} &= \frac{z + 1}{2z - 1} + 1, \\ \frac{dz}{dx} &= \frac{3z}{2z - 1}, \\ \left(\frac{2}{3} - \frac{1}{3z}\right) dz &= 1 dx, \\ \frac{2}{3} z - \frac{1}{3} \ln |z| &= x + C, \quad C \in \mathbb{R}, \end{aligned}$$

or

$$\frac{2}{3} z - \frac{1}{3} \ln |Cz| = x, \quad C \neq 0.$$

Now, we must get back to the original variable  $y$  in one of these forms. The general solution can be written as

$$\frac{2}{3} x + \frac{2}{3} y - \frac{1}{3} \ln |x + y| = x + C, \quad C \in \mathbb{R},$$

i. e.,

$$x - 2y + \ln |x + y| = C, \quad C \in \mathbb{R}.$$

At the same time, we have the singular solution  $y = -x$ , which follows from the constraint  $z \neq 0$  of the operations we have made (we have divided by the value  $3z$ ).  $\square$

**8.J.6.** Solve the differential equation

$$xy' + y \ln x = y \ln y.$$

**Solution.** Using the substitution  $u = y/x$ , every homogeneous differential equation  $y' = f(y/x)$  can be transformed to an equation (with separated variables)

$$u' = \frac{1}{x} (f(u) - u), \quad \text{i. e.} \quad u'x + u = f(u).$$

The name of this differential equation is comes from the following definition. A function  $f$  of two variables is called homogeneous of degree  $k$  iff  $f(tx, ty) = t^k f(x, y)$ . Then, a differential equation of the form

$$P(x, y) dx + Q(x, y) dy = 0$$

or the natural decrease of stock population as a result of sending some part of it to slaughterhouses), an equation can be used with a constant  $s$  on the right-hand side.

$$(2) \quad y'(t) = r \cdot y(t) + s.$$

The solution of this equation is the function

$$y(t) = C e^{rt} - \frac{s}{r}.$$

It is a straightforward matter to produce this solution when it is realized that the set of all solutions of the equation (1) is a one-dimensional vector space, while the solutions of the equation (2) are obtained by adding any one of its solutions to the solutions of the previous equation. The constant solution  $y(t) = k$  for  $k = -\frac{s}{r}$  is easily found.

Similarly, in paragraph 1.4.1, the *logistic model* of population growth was created. Based on the assumption that the ratio of the change of the population size  $p(n + 1) - p(n)$  and its size  $p(n)$  is affine with respect to the population size itself. The model behaves similar as the Malthusian one for small values of the population size and to cease growing when reaching a limit value  $K$ . Now, the same relation for the continuous model can be formulated for a population  $p(t)$  dependent on time  $t$  by the equality

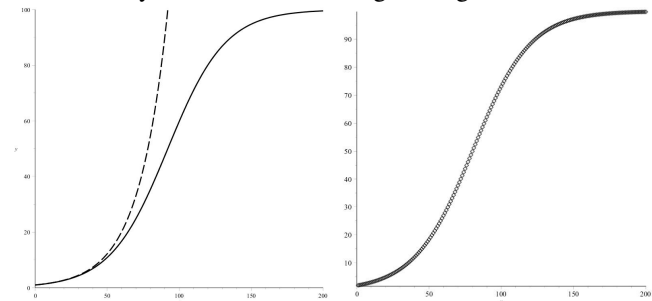
$$(3) \quad p'(t) = p(t) \left( -\frac{r}{K} p(t) + r \right).$$

At the value  $p(t) = K$  for a (large) constant  $K$ , the instantaneous increment of the function  $p$  is zero, while for  $p(t) > 0$  near zero, the ratio of the rate of increment of the population and its size is close to  $r$ , which is the (small) number expressing the rate of increment of the population in good conditions (e.g. 0.05 would again mean immediate growth by 5%).

It is not easy to solve such an equation without knowing any theory (although this type of equations will be dealt with in a moment). However, as an exercise on differentiation, it is easily verified that the following function is a solution for every constant  $C$ :

$$p(t) = \frac{K}{1 + CK e^{-rt}}.$$

For the continuous and discrete versions of the logistic models, the values  $K = 100$ ,  $r = 0.05$ , and  $C = 1$  in the left hand illustration are chosen. The same 1.4.1 result occurs in the right hand illustration (i.e. with  $a = 1.05$  and  $p_1 = 1$ , as expected). The choice  $C = 1$  yields  $p(0) = K/(1 + K)$  which is very close to 1 if  $K$  is large enough.



is a homogeneous differential equation iff the functions  $P$  and  $Q$  are homogeneous of the same degree  $k$ .

For instance, we can discover that the given equation

$$x dy + (y \ln x - y \ln y) dx = 0$$

is homogeneous. Of course, it is not difficult to write it explicitly in the form

$$y' = \frac{y}{x} \ln \frac{y}{x}.$$

The substitution  $u = y/x$  then leads to

$$\begin{aligned} u'x + u &= u \ln u, \\ \frac{du}{dx} x &= u (\ln u - 1), \\ \frac{du}{u (\ln u - 1)} &= \frac{dx}{x}, \end{aligned}$$

where  $u (\ln u - 1) \neq 0$ . Using another substitution, namely  $t = \ln u - 1$ , we can integrate

$$\begin{aligned} \int \frac{du}{u (\ln u - 1)} &= \int \frac{dx}{x}, \\ \int \frac{dt}{t} &= \int \frac{dx}{x}, \\ \ln |t| &= \ln |x| + \ln |C|, \quad C \neq 0, \\ \ln |\ln u - 1| &= \ln |Cx|, \quad C \neq 0, \\ \ln u - 1 &= Cx, \quad C \neq 0, \\ \ln \frac{y}{x} &= Cx + 1, \quad C \neq 0, \\ y &= xe^{Cx+1}, \quad C \neq 0. \end{aligned}$$

The excluded cases  $u = 0$  and  $\ln u = 1$  do not lead to two more solutions since  $u = 0$  implies  $y = 0$ , which cannot be put into the original equation. On the other hand,  $\ln u = 1$  gives  $y/x = e$ , and the function  $y = ex$  is clearly a solution. Therefore, the general solution is

$$y = xe^{Cx+1}, \quad C \in \mathbb{R}.$$

□

**8.J.7.** Compute

$$y' = -\frac{4x+3y+1}{3x+2y+1}.$$

**Solution.** In general, we are able to solve every equation of the form

$$(1) \quad y' = f\left(\frac{ax + by + c}{Ax + By + C}\right).$$

If the system of linear equations

$$(2) \quad ax + by + c = 0, \quad Ax + By + C = 0$$

has a unique solution  $x_0, y_0$ , then the substitution  $u = x - x_0$ ,  $v = y - y_0$  transforms the equation (1) to a homogeneous equation

$$\frac{dv}{du} = f\left(\frac{au + bv}{Au + Bv}\right).$$

In particular, both versions of this logistic model yield quite similar results. For example, the left hand illustration also contains the dashed line of the graph of the solution of the equation (1) with the same constant  $r$  and initial condition (i.e. the Malthusian model of growth).

**8.3.2. First-order differential equations.** By an (ordinary) first-order differential equation, is usually meant the relation between the derivative  $y'(t)$  of a function with respect to the variable  $t$ , its value  $y(t)$ , and the variable itself, which can be written in terms of some real-valued function  $F : \mathbb{R}^3 \rightarrow \mathbb{R}$  as the equality



$$F(y'(t), y(t), t) = 0.$$

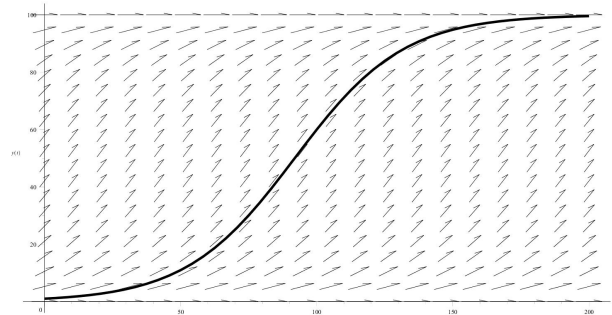
This equation resembles the implicitly defined functions  $y(t)$ ; however, this time there is a dependency on the derivative of the function  $y(t)$ . We also often suppress the dependence of  $y = y(t)$  on the other variable  $t$  and write  $F(y', y, t) = 0$  instead.

If the implicit equation is solved at least explicitly with regard to the derivative, i.e.

$$y' = f(t, y)$$

for some function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ , it is clear graphically what this equation defines. For every value  $(t, y)$  in the plane, the arrow corresponding to the vector  $(1, f(t, y))$ , can be considered. That is the velocity with which the point of the graph of the solution moves through the plane, depending on the free parameter  $t$ .

For instance, the equation (3) in the previous subsection determines the following: (illustrating the solution for the initial condition as above).



Such illustrations should invoke the idea that differential equations define a “flow” in the plane, and each choice of the initial value  $(t_0, y(t_0))$  should correspond to a unique flow-line expressing the movement of the initial point in the time  $t$ . It can be anticipated intuitively that for reasonably behaved functions  $f(t, y)$  in the equations  $y' = f(t, y)$ , there is a unique solution for all initial conditions.



If the system (2) has no solution or has infinitely many solutions, the substitution  $z = ax + by$  transforms the equation (1) to an equation with separated variables (often, the original equation is already such).

In this problem, the corresponding system of equations

$$4x + 3y + 1 = 0, \quad 3x + 2y + 1 = 0$$

has a unique solution  $x_0 = -1, y_0 = 1$ . The substitution  $u = x+1, v = y-1$  then leads to the homogeneous equation

$$\frac{dv}{du} = -\frac{4u+3v}{3u+2v},$$

which can be solved by further substitution  $z = v/u$ . We thus obtain

$$\begin{aligned} z'u + z &= -\frac{4 + 3z}{3 + 2z}, \\ \frac{dz}{du} u &= -\frac{2z^2 + 6z + 4}{3 + 2z}, \\ \frac{2z + 3}{2z^2 + 6z + 4} dz &= -\frac{du}{u} \end{aligned}$$

provided  $z^2 + 3z + 2 \neq 0$ . Integrating, we get

$$\begin{aligned} \frac{1}{2} \ln |z^2 + 3z + 2| &= -\ln |u| + \ln |C|, \quad C \neq 0, \\ \frac{1}{2} \ln |(z^2 + 3z + 2) u^2| &= \ln |C|, \quad C \neq 0, \\ \ln |(z^2 + 3z + 2) u^2| &= \ln C^2, \quad C \neq 0, \\ (z^2 + 3z + 2) u^2 &= \pm C^2, \quad C \neq 0. \end{aligned}$$

We thus have

$$(z^2 + 3z + 2) u^2 = D, \quad D \neq 0$$

and returning to the original variables,

$$\begin{aligned} \left(\frac{v^2}{u^2} + 3\frac{v}{u} + 2\right) u^2 &= D, \quad D \neq 0, \\ v^2 + 3vu + 2u^2 &= D, \quad D \neq 0, \\ (y - 1)^2 + 3(y - 1)(x + 1) + 2(x + 1)^2 &= D, \quad D \neq 0. \end{aligned}$$

Making simple rearrangements, the general solution can be expressed as

$$(x + y)(2x + y + 1) = D, \quad D \neq 0.$$

Now, let us return to the condition  $z^2 + 3z + 2 \neq 0$ . It follows from  $z^2 + 3z + 2 = 0$  that  $z = -1$  or  $z = -2$ , i. e.,  $v = -u$  or  $v = -2u$ . For  $v = -u$ , we have  $x = u - 1$  and  $y = v + 1 = -u + 1$ , which means that  $y = -x$ . Similarly, for  $v = -2u$ , we have  $y = -2u + 1$ , hence  $y = -2x - 1$ . However, both functions  $y = -x, y = -2x - 1$  satisfy the original differential equations and are included in the general solution for the choice  $D = 0$ . Therefore, every solution is known from the implicit form

$$(x + y)(2x + y + 1) = D, \quad D \in \mathbb{R}.$$

**8.3.3. Integration of differential equations.** Before examining the conditions for existence and uniqueness of the solutions, we present a truly elementary method for finding the solutions. The idea, mentioned briefly already in 6.2.21 on page 552, is to transform the problem to ordinary integration, which usually leads to an implicit description of the solution.



EQUATIONS WITH SEPARATED VARIABLES

Consider a differential equation in the form

$$(1) \quad y' = f(t) \cdot g(y)$$

for two continuous functions of a real variable,  $f$  and  $g$ .

The solution of this equation can be obtained by integration, finding the antiderivatives

$$G(y) = \int \frac{dy}{g(y)}, \quad F(t) = \int f(t) dt.$$

This procedure reliably finds solutions  $y(t)$  which satisfy  $g(y(t)) \neq 0$ , given implicitly by the formula

$$(2) \quad F(t) + C = G(y)$$

with an arbitrary constant  $C$ .

Differentiating the latter equation (2) using the chain rule for the composite function  $G(y(t))$  leads to  $\frac{1}{g(y)} y'(t) = f(t)$ , as required.

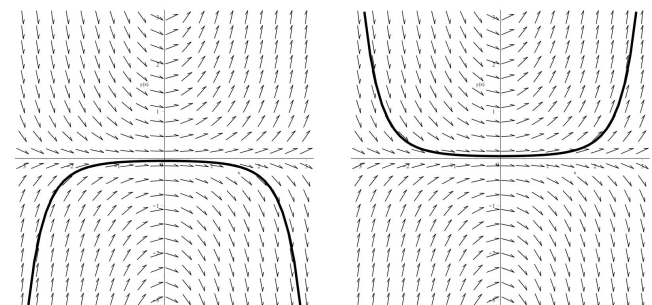
As an example, find the solution of the equation

$$y' = t y.$$

Direct calculation gives  $\ln |y(t)| = \frac{1}{2} t^2 + C$  with arbitrary constant  $C$ . Hence it looks (at least for positive values of  $y$ ) as

$$y(t) = e^{\frac{1}{2} t^2 + C} = D e^{\frac{1}{2} t^2},$$

where  $D$  is an arbitrary positive constant. It is helpful to examine the resulting formula and signs thoroughly. The constant solution  $y(t) = 0$  also satisfies the equation. For negative values of  $y$ , the same solution can be used with negative constants  $D$ . In fact, the constant  $D$  can be arbitrary, and a solution is found satisfying any initial value.



The illustration shows two solutions which demonstrate the instability of the equation with regard to the initial values: For every  $t_0$ , if we change a small  $y_0$  from a negative value to a positive one, then the behaviour of the resulting solution changes dramatically. Notice the constant solution  $y(t) = 0$ , which satisfies the initial condition  $y(t_0) = 0$ .

□

**8.J.8.** Find the general solution of the differential equation

$$(x^2 + y^2) dx - 2xy dy = 0.$$

**Solution.** For  $y \neq 0$ , simple rearrangements lead to

$$y' = \frac{x^2 + y^2}{2xy} = \frac{1 + (\frac{y}{x})^2}{2\frac{y}{x}}.$$

Using the substitution  $u = y/x$ , we get to the equation

$$u'x + u = \frac{1+u^2}{2u}.$$

For  $u \neq \pm 1$  and  $D = -1/C$ , we have

$$\begin{aligned} \frac{du}{dx} x &= \frac{1 + u^2 - 2u^2}{2u}, \\ \frac{2u}{1 - u^2} du &= \frac{dx}{x}, \\ -\ln |1 - u^2| &= \ln |x| + \ln |C|, \quad C \neq 0, \\ \ln \frac{1}{|1 - u^2|} &= \ln |Cx|, \quad C \neq 0, \\ \frac{1}{1 - u^2} &= Cx, \quad C \neq 0, \\ 1 &= Cx \left(1 - \frac{y^2}{x^2}\right), \quad C \neq 0, \\ -\frac{D}{x} &= 1 - \frac{y^2}{x^2}, \quad D \neq 0, \\ -Dx &= x^2 - y^2, \quad D \neq 0. \end{aligned}$$

The condition  $u = \pm 1$  corresponds to  $y = \pm x$ . While  $y \equiv 0$  is not a solution, both the functions  $y = x$  and  $y = -x$  are solutions and can be obtained by the choice  $D = 0$ . The general solution is thus

$$y^2 = x^2 + Dx, \quad D \in \mathbb{R}. \quad \square$$

**8.J.9.** Solve

$$y' = x - \frac{2y}{x^2 - 1}.$$

**Solution.** The given equation is of the form  $y' = a(x)y + b(x)$ , i. e., a non-homogeneous linear differential equation (the function  $b$  is not identically equal to zero). The general solution of such an equation can be obtained using the method of integration factor (the non-homogeneous equation is multiplied by the expression  $e^{-\int a(x) dx}$ ) or the method of variable separation (the integration constant that arises in the solution of the corresponding homogeneous equations is considered to be a function in the variable  $x$ ). We will illustrate both of these methods on this problem.

As for the former method, we multiply the original equation by the expression

$$e^{\int \frac{2}{x^2-1} dx} = e^{\ln \left| \frac{x-1}{x+1} \right|} = \frac{x-1}{x+1},$$

where the corresponding integral is understood to stand for any antiderivative and where any non-zero multiple of the obtained function can be considered (that is why we could remove the absolute value). Thus, consider the equation

$$y' \frac{x-1}{x+1} + \frac{2y}{(x+1)^2} = \frac{x(x-1)}{x+1}.$$

The core of the method of integration factor is that fact that the expression on the left-hand side is the derivative of  $y \frac{x-1}{x+1}$ . Integrating this leads to

Using separation of variables, the non-linear equation is easily solved from the previous paragraph which describes the logistic population model. Try this as an exercise.

**8.3.4. First order linear equations.** In the first chapter, we



paid much attention to linear difference equations. Their general solution was determined in paragraph 1.2.2 on page 11. Although it is clear beforehand that it is a one-dimensional affine space of sequences, it is a hardly transparent sum, because all the changing coefficients need to be taken into account.

Consequently this can be used as a source of inspiration for the following construction of the solution of a general first-order linear equation

$$(1) \quad y' = a(t)y + b(t)$$

with continuous coefficients  $a(t)$  and  $b(t)$ .

First, find the solution of the homogeneous equation  $y' = a(t)y$ . This can be computed easily by separation of variables, obtaining the solution with  $y(s) = y_0$ ,

$$y(t) = y_0 F(t, t_0), \quad F(t, s) = e^{\int_s^t a(x) dx}.$$

In the case of difference equations, the solution of the general non-homogeneous equation was “guessed”. Then it was proved by induction that it was correct. It is even simpler now, as it suffices to differentiate the correct solution to verify the statement, once we are told what the right result is:

#### THE SOLUTION OF FIRST-ORDER LINEAR EQUATIONS

The solution of the equation (1) with initial values  $y(t_0) = y_0$  is (locally in a neighbourhood of  $t_0$ ) given by the formula

$$y(t) = y_0 F(t, t_0) + \int_{t_0}^t F(t, s) b(s) ds,$$

where  $F(t, s) = e^{\int_s^t a(x) dx}$ .

Verify the correctness of the solution by yourselves (pay proper attention to the differentiation of the integral where  $t$  is both in the upper bound and a free parameter in the integrand, cf. 6.3.13).

In fact, there is the general method called *variation of constants* which directly yields this solution, see e.g. the problem 8.J.9. It consists in taking the solution for the homogenous equation in the form  $y(t) = cF(t, t_0)$  and consider instead an ansatz for a solution to the non-homogeneous equation in the form  $y(t) = c(t)F(t, t_0)$  with an unknown function  $c(t)$ . Differentiating yields the equation  $c' = e^{-\int_{t_0}^t a(x) dx} b(t)$  and integrating this leads to  $c(t) = \int_{t_0}^t e^{\int_s^{t_0} a(x) dx} b(s) ds$ , i.e.  $y(t) = c(t) e^{\int_{t_0}^t a(x) dx}$  as in the above formula. Check the details!

Notice also the similarity to the solution for the equations with constant coefficients explicitly computed in the form of convolution in 7.B.13 on the page 663, which could serve as inspiration, too.

As an example, the equation

$$y' = 1 - xy,$$

$$y \frac{x-1}{x+1} = \int \frac{x(x-1)}{x+1} dx = \frac{x^2}{2} - 2x + 2 \ln|x+1| + C, \quad C \in \mathbb{R}.$$

Therefore, the solutions are the functions

$$y = \frac{x+1}{x-1} \left( \frac{x^2}{2} - 2x + 2 \ln|x+1| + C \right), \quad C \in \mathbb{R}.$$

As for the latter method, we first solve the corresponding homogeneous equation

$$y' = -\frac{2y}{x^2-1},$$

which is an equation with separated variables. We have

$$\begin{aligned} \frac{dy}{dx} &= -\frac{2y}{x^2-1}, \\ \frac{dy}{y} &= -\frac{2}{x^2-1} dx, \end{aligned}$$

$$\ln|y| = -\ln|x-1| + \ln|x+1| + \ln|C|, \quad C \neq 0,$$

$$\ln|y| = \ln \left| C \frac{x+1}{x-1} \right|, \quad C \neq 0,$$

$$y = C \frac{x+1}{x-1}, \quad C \neq 0,$$

where we had to exclude the case  $y = 0$ . However, the function  $y \equiv 0$  is always a solution of a homogeneous linear differential equation, and it can be included in the general solution. Therefore, the general solution of the corresponding homogeneous equation is

$$y = \frac{C(x+1)}{x-1}, \quad C \in \mathbb{R}.$$

Now, we will consider the constant  $C$  to be a function  $C(x)$ . Differentiating leads to

$$y' = \frac{C'(x)(x+1)(x-1) + C(x)(x-1) - C(x)(x+1)}{(x-1)^2}.$$

Substituting this into the original equation, we get

$$\frac{C'(x)(x+1)(x-1) + C(x)(x-1) - C(x)(x+1)}{(x-1)^2} = x - \frac{2C(x)(x+1)}{(x-1)(x^2-1)}.$$

It follows that

$$C'(x) = \frac{x(x-1)}{x+1},$$

i. e.,

$$C(x) = \int \frac{x(x-1)}{x+1} dx,$$

$$C(x) = \frac{x^2}{2} - 2x + 2 \ln|x+1| + C, \quad C \in \mathbb{R}.$$

Now, it suffices to substitute:

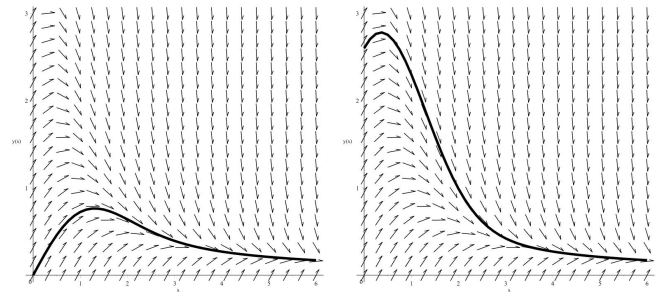
$$\begin{aligned} y &= C(x) \frac{x+1}{x-1} = \\ \frac{x+1}{x-1} \left( \frac{x^2}{2} - 2x + 2 \ln|x+1| + C \right), \quad C \in \mathbb{R}. \end{aligned}$$

We can see that the result we have obtained here is of the same form as in the former case. This should not be surprising as the differences between the two methods are insignificant and the computed integrals are the same.

Finally, we can notice that the solution  $y$  of an equation  $y' = a(x)y$  can be found in the same way for any continuous function  $a$ . We thus always have

$$y = Ce^{\int a(x) dx}, \quad C \in \mathbb{R}.$$

can be solved directly (although the so-called error function appears when integrating the particular solution and this cannot be expressed via elementary functions), this time encountering stable behaviour, visible in the following illustration.



**8.3.5. Transformation of coordinates.** The illustrations suggest that differential equations can be perceived as geometric objects (the “directional field of the arrows”), so the solution can be found by conveniently chosen coordinates. We return to this point of view later. Here are three simple examples of typical tricks as seen from the explicit form of the equations in coordinates.

We begin with *homogeneous equations* of the form

$$y' = f\left(\frac{y}{t}\right).$$

Considering the transformation  $z = z(t) = \frac{y}{t}$  and assuming that  $t \neq 0$ , then by the chain rule,

$$z' = \frac{1}{t^2}(ty' - y) = \frac{1}{t}(f(z) - z),$$

which is an equation with separated variables.

Other examples are the *Bernoulli differential equations*, which are of the form

$$y' = f(t)y + g(t)y^n,$$

where  $n \neq 0, 1$ . The choice of the transformation  $z = y^{1-n}$  leads to the equation

$$\begin{aligned} z' &= (1-n)y^{-n}(f(t)y + g(t)y^n) \\ &= (1-n)f(t)z + (1-n)g(t), \end{aligned}$$

which is a linear equation, easily integrated.

We conclude with the extraordinarily important *Riccati equation*. It is a form of the Bernoulli equation with  $n = 2$ , extended by an absolute term

$$y' = f(t)y + g(t)y^2 + h(t).$$

This equation can also be transformed to a linear equation provided that a particular solution  $x = x(t)$  can be guessed. Then, use the transformation

$$z = \frac{1}{y-x}.$$

Verify by yourselves that this transformation leads to the equation

$$z' = -(f(t) + 2x(t)g(t))z - g(t).$$

Similarly, the solution of an equation  $y' = a(x)y + b(x)$  with an initial condition  $y(x_0) = y_0$  can be determined explicitly as (provided the coefficients, i. e. the functions  $a$  and  $b$ , are continuous)

$$y = e^{\int_{x_0}^x a(t) dt} \left( y_0 + \int_{x_0}^x b(t) e^{-\int_{x_0}^t a(s) ds} dt \right).$$

Let us remark that the linear equation has no singular solution, and the general solution contains a  $C \in \mathbb{R}$ .  $\square$

**8.J.10.** Solve the linear equation

$$(y' + 2xy) e^{x^2} = \cos x.$$

**Solution.** If we used the method of integration factor, we would only rewrite the equation trivially since it is already of the desired form – the expression on the left-hand side is the derivative of  $y e^{x^2}$ . Thus, we can immediately calculate

$$\begin{aligned} (y e^{x^2})' &= \cos x, \\ y e^{x^2} &= \int \cos x dx, \\ y e^{x^2} &= \sin x + C, \quad C \in \mathbb{R}, \\ y &= e^{-x^2} (\sin x + C), \quad C \in \mathbb{R}. \end{aligned}$$

$\square$

**8.J.11.** Find all non-zero solutions of the Bernoulli equation

$$y' - \frac{y}{x} = 3xy^2.$$

**Solution.** The Bernoulli equation

$$y' = a(x)y + b(x)y^r, \quad r \neq 0, r \neq 1, r \in \mathbb{R}$$

can be solved by first dividing by the term  $y^r$  and then using the substitution  $u = y^{1-r}$ , which leads to the linear differential equation

$$u' = (1 - r) [a(x)u + b(x)].$$

In this very problem, the substitution  $u = y^{1-2} = 1/y$  gives

$$u' + \frac{u}{x} = -3x.$$

Similarly to the previous exercise, we have

$$u = e^{-\ln|x|} \left[ \int -3x e^{\ln|x|} dx \right],$$

where  $\ln|x|$  was obtained as an (arbitrary) antiderivative to  $1/x$ . Further,

$$\begin{aligned} u &= e^{\ln \frac{1}{|x|}} \left[ \int -3x e^{\ln|x|} dx \right], \\ u &= \frac{1}{|x|} \left[ \int -3x |x| dx \right]. \end{aligned}$$

The absolute value can be replaced with a sign that can be canceled, i. e., it suffices to consider

$$u = \frac{1}{x} \left[ \int -3x^2 dx \right] = \frac{1}{x} [-x^3 + C], \quad C \in \mathbb{R}.$$

Returning to the original variable, we get

$$y = \frac{1}{u} = \frac{x}{C - x^3}, \quad C \in \mathbb{R}.$$

The excluded case  $y \equiv 0$  is a singular solution (which, of course, is true for every Bernoulli equation with  $r$  positive).  $\square$

As seen in the case of integration of functions (the simplest type of equations with separated variables), the equations usually do not have a solution expressible explicitly in terms of elementary functions.

As with standard engineering tables of values of special functions, books listing the solutions of basic equations are compiled as well.<sup>7</sup> Today, the wisdom concealed in them is essentially transferred to software systems like Maple or Mathematica. Here, any task about ordinary differential equations can be assigned, with results obtained in surprisingly many cases. Yet, explicit solutions are not possible for most problems.

**8.3.6. Existence and uniqueness.** The way out of this is numerical methods, which try only to approximate the solutions. However, to be able to use them, good theoretical starting points are still needed regarding existence, uniqueness, and stability of the solutions.



We begin with the *Picard–Lindelöf theorem*:

EXISTENCE AND UNIQUENESS OF THE SOLUTIONS OF ODES

**Theorem.** Consider a function  $f(t, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$  with continuous partial derivatives on an open set  $U$ . Then for every point  $(t_0, y_0) \in U \subset \mathbb{R}^2$ , there exists the maximal interval  $I = (t_0 - a, t_0 + b)$ , with positive  $a, b \in \mathbb{R}$ , and the unique function  $y(t) : I \rightarrow \mathbb{R}$  which is the solution of the equation  $y' = f(t, y)$  on the interval  $I$ .

**PROOF.** If a differentiable function  $y(t)$  is a solution of an equation satisfying the initial condition  $y(t_0) = y_0$ , then it also satisfies the equation

$$y(t) = y_0 + \int_{t_0}^t y'(s) ds = y_0 + \int_{t_0}^t f(s, y(s)) ds,$$

where the Riemann integrals exist due to the continuity of  $f$  and hence also  $y'$ . However, the right-hand side of this expression is the integral operator

$$L(y)(t) = y_0 + \int_{t_0}^t f(s, y(s)) ds$$

acting on functions  $y$ . Solving first-order differential equations, is equivalent to finding fixed points for this operator  $L$ , that is, to find a function  $y = y(t)$  satisfying  $L(y) = y$ .

On the other hand, if a Riemann-integrable function  $y(t)$  is a fixed point of the operator  $L$ , then it immediately follows from the fundamental theorem of calculus that  $y(t)$  satisfies the given differential equation, including the initial conditions.

<sup>7</sup>For example, the famous book *Differentialgleichungen reeller Funktionen*, Akademische Verlagsgesellschaft, Leipzig 1930, by E. Kamke, a German mathematician, contains many hundreds of solved equations. They appeared in many editions in the last century.

**8.J.12.** Interchanging the variables, solve the equation

$$y dx - (x + y^2 \sin y) dy = 0.$$

**Solution.** When the variable  $x$  occurs only in the first power in the differential equation and  $y$  occurs in the arguments of elementary functions, we can apply the so-called method of variable interchange, when we look for the solution as for a function  $x$  of the independent variable  $y$ .

First, we write the equation explicitly:

$$y' = \frac{y}{x + y^2 \sin y}.$$

This equation is not of any of the previous types, so we rewrite it as follows:

$$\begin{aligned} \frac{dy}{dx} &= \frac{y}{x + y^2 \sin y}, \\ \frac{dx}{dy} &= \left( \frac{y}{x + y^2 \sin y} \right)^{-1} = \frac{x}{y} + y \sin y, \\ x' &= \frac{1}{y} x + y \sin y. \end{aligned}$$

We have thus obtained a linear differential equation. Now, we can easily compute its general solution

$$x = -y \cos y + Cy, \quad C \in \mathbb{R}.$$

□

Further problems concerning first-order differential equations can be found on page 795.

### K. Practical problems leading to differential equations

**8.K.1.** A water purification plant with volume  $2000 \text{ m}^3$  was contaminated with lead which is spread in the water with density  $10 \text{ g/m}^3$ . Water is flowing in and out of the basin at  $2 \text{ m}^3/\text{s}$ . In what time does the amount of lead in the basin decrease below  $10 \text{ } \mu\text{g/m}^3$  (which is the hygienic norm for the amount of lead in drinkable water by a regulation of the European Community) provided the water keeps being mixed uniformly?

**Solution.** Let us denote the water's volume in the basin by  $V \text{ (m}^3\text{)}$ , the speed of the water's flow by  $v \text{ (m}^3/\text{s}\text{)}$ . In an infinitesimal (infinitely small) time unit  $dt$ ,  $\frac{m}{V} \cdot v dt$  grams of lead runs out of the basin, so we can construct the differential equation

$$dm = -\frac{m}{V} \cdot v dt$$

for the change of the lead's mass in the basin. Separating the variables, we get the equation

$$\frac{dm}{m} = -\frac{v}{V} dt.$$

Integration both sides of the equation and getting rid of the logarithms, we get the solution in the form  $m(t) = m_0 e^{-\frac{v}{V}t}$ , where  $m_0$  is the lead's mass at time  $t = 0$ . Substituting the concrete values, we find out that  $t \doteq 6 \text{ h } 35 \text{ min}$ . □



It is easy to estimate how much the values  $L(y)$  and  $L(z)$  differ for various functions  $y(t)$  and  $z(t)$ . Since both partial derivatives of  $f$  are continuous,  $f$  is itself locally Lipschitz. This means that restricting the values  $(t, y)$  to a neighbourhood  $U$  of the point  $(t_0, y_0)$  with compact closure, there is the estimate

$$|f(t, y) - f(t, z)| \leq C|y - z|,$$

with some constant  $C$  depending only on  $U$ . This immediately leads to the following bound (for the sake of simplicity,  $t \geq t_0$ , but the final conclusion works for  $t < t_0$  the same way)

$$\begin{aligned} |(L(y) - L(z))(t)| &= \left| \int_{t_0}^t f(s, y(s)) - f(s, z(s)) ds \right| \\ &\leq \int_{t_0}^t |f(s, y(s)) - f(s, z(s))| ds \\ &\leq C \int_{t_0}^t |y(s) - z(s)| ds \\ &\leq C \left( \max_{t_0 \leq s \leq t} |y(s) - z(s)| \right) |t - t_0| \\ &= D \left( \max_{t_0 \leq s \leq t} |y(s) - z(s)| \right), \end{aligned}$$

where the constant  $D$  comes from substituting the maximum of  $|t - t_0|$  on  $U$ .

If the operator  $L$  is viewed as an operator on a metric space of continuous functions on a compact interval with the max norm, this yields

$$\|L(y) - L(z)\| \leq D \|y - z\|.$$

Some further restrictions on the choice of  $U$  and the considered functions  $y$  and  $z$  are required, in order to make the constant  $D$  smaller than one. Then the Banach fixed point theorem, based on the notion of a contraction, can be applied. See 7.3.9 on the page 676. At the same time, the operator must leave the chosen space of functions  $y$  invariant, i.e. the images  $L(y)$  are also there.



To begin, choose  $\varepsilon > 0$  and  $\delta > 0$ , both small enough so that  $[t_0 - \delta, t_0 + \delta] \times [y_0 - \varepsilon, y_0 + \varepsilon] = V \subset U$ , and consider only those functions  $y(t)$  which satisfy for  $J = [t_0 - \delta, t_0 + \delta]$  the estimate  $\max_{t \in J} |y(t) - y_0| < \varepsilon$ . The uniform continuity of  $f(t, y)$  on  $V$  ensures that fixing  $\varepsilon$  and further shrinking  $\delta$ , implies

$$\max_{t \in J} |L(y)(t) - y_0| < \varepsilon.$$

Finally, the above estimate for  $\|L(y) - L(z)\|$  shows that if  $\delta$  is decreased sufficiently further, then the latter constant  $D$  becomes smaller than one, as required for a contraction. At the same time,  $L$  maps the above space of functions into itself.

However, for the assumptions of the Banach contraction theorem, which guarantees the uniquely determined fixed point, completeness of the space  $X$  of functions on which the operator  $L$  works is needed.

Since the mapping  $f(t, y)$  is continuous, there follows a uniform bound for all of the functions  $y(t)$  considered above

**8.K.2.** The speed of transmission of a message in a population consisting of  $P$  people is directly proportional to the number of people who have not heard the message yet. Determine the function  $f$  which describes the dependency of the number of people who *have* heard the message on time. Is it appropriate to use this model of message transmission for small or large values of  $P$ ?

**Solution.** We construct a differential equation for  $f$ . The speed of the transmission  $\frac{df}{dt} = f'(t)$  should be directly proportional to the number of people who have not heard of it, i. e. the value  $P - f(t)$ . Altogether,

$$\frac{df}{dt} = k(P - f(t)).$$

Separating the variables and introducing a constant  $K$  (the number of people who know the message at time  $t = 0$  must be  $P - K$ ), we get the solution

$$f(t) = P - Ke^{-kt},$$

where  $k$  is a positive real constant.

Apparently, this model makes sense for large values of  $P$  only. □

**8.K.3.** The speed at which an epidemic spreads in a given closed population consisting of  $P$  people is directly proportional to the product of the number of people who have been infected and the number of people who have not. Determine the function  $f(t)$  describing the number of infected people in time.

**Solution.** Just like in the previous problem, we construct a differential equation:

$$\frac{df}{dt} = k \cdot f(t) (P - f(t)).$$

Again, separating the variables and introducing suitable constants  $K$  and  $L$ , we obtain

$$f(t) = \frac{K}{1 + Le^{-Kkt}}.$$

□

**8.K.4.** The speed at which a given isotope of a given chemical element decays is directly proportional to the amount of the given isotope. The half-life of the isotope of plutonium  ${}_{94}^{239}\text{Pu}$  is 24,100 years. In what time does a hundredth of a nuclear bomb whose active component is the mentioned isotope disappear?

**Solution.** Denoting the amount of plutonium by  $m$ , we can build a differential equation for the rate of the decay:

$$\frac{dm}{dt} = k \cdot m,$$

where  $k$  is an unknown constant. The solution is thus the function  $m(t) = m_0 e^{-kt}$ . Substituting into the equation for half-life ( $e^{-kt} = \frac{1}{2}$ ), we get the constant  $k \doteq 2.88 \cdot 10^5$ . The wanted time is then approximately 349 years. □

and the values  $t > s$  in their domain:

$$|L(y)(t) - L(y)(s)| \leq \int_s^t |f(s, y(s))| ds \leq A |t - s|$$

with a universal constant  $A > 0$ . Besides the conditions mentioned above, there is a restriction to the subset of all equicontinuous functions in the sense of the Definition 7.3.15. According to the Arzelà-Ascoli Theorem proved in the same paragraph at the page 683, this set of continuous functions is already compact, hence it is a complete set of continuous functions on the interval.

Therefore, there exists a unique fixed point  $y(t)$  of this contraction  $L$  by the Theorem 7.3.9. This is the solution of the equation.

It remains to show the existence of a maximal interval  $I = (t_0 - a, t_0 + b)$ . Suppose that a solution  $y(t)$  is found on an interval  $(t_0, t_1)$ , and, at the same time, the one-sided limit  $y_1 = \lim_{t \rightarrow t_1^-} y(t)$  exists and is finite.

It follows from the already proven result that there exists a solution with this initial condition  $(t_1, y_1)$ , in some neighbourhood of the point  $t_1$ . Clearly, it must coincide with the discussed solution  $y(t)$  on the left-hand side of  $t_1$ . Therefore, the solution  $y(t)$  can be extended on the right-hand side of  $t_1$ .

There are only two possibilities when the extension of the solution behind  $t_1$  does not exist: either there is no finite left limit  $y(t)$  at  $t_1$ , or the limit  $y_1$  exists, yet the point  $(t_1, y_1)$  is on the boundary of the domain of the function  $f$ . In both cases, the maximal extension of the solution to the right of  $t_0$  is found.

The argumentation for the maximal solution left of  $t_0$  is analogous. □

**8.3.7. Iterative approximations of solutions.** The proof of the previous theorem can be reformulated as an iterative procedure which provides approximate solutions using step-by-step integration. Moreover, an explicit estimate for the constant  $C$  from the proof yields bounds for the errors.



Think this out as an exercise (see the proof of Banach fixed-point theorem in paragraph 7.3.9). It can then be shown easily and directly that it is a uniformly convergent sequence of continuous functions, so the limit is again a continuous function (without invoking the complicated theorems from the seventh chapter).

**8.K.5.** The acceleration of an object falling in a constant gravitational field with a certain resistance of the environment is given by the formula

$$\frac{dv}{dt} = g - kv,$$

where  $k$  is a constant which expresses the resistance of the environment. An object was dropped in a gravitational field with  $g = 10 \text{ ms}^{-2}$  at the initial speed of  $5 \text{ ms}^{-1}$ , the resistance constant is  $k = 0.5 \text{ s}^{-1}$ . What will the speed of the object be in three seconds?

**Solution.**

$$v = \frac{g}{k} - \left(\frac{g}{k} - v_0\right) e^{-kt},$$

$v(3) = 20 - 15e^{-\frac{3}{2}} \text{ ms}^{-1}$  after substitution. □

**8.K.6.** The rate of increase of a population of a certain type of bug is indirectly proportional to its size. At time  $t = 0$ , the population had 100 bugs. In a month, the population doubled. What will the size of the population be in two months?

**Solution.** Let us consider a continuous approximation of the number of bugs, and let their amount be denoted by  $P$ . Then, we can build the following equation:

$$\frac{dP}{dt} = \frac{k}{P},$$

$P = \sqrt{Kt + c}$ . Substituting the given values, we get  $P(2) = \sqrt{7} \cdot 100$ , which is an estimate of the actual number of bugs. □

**8.K.7.** Find the equation of the curve with the following properties: It lies in the first quadrant, goes through the point  $[1, 3/4]$ , and its tangent at any point marks on the positive half-axis  $y$  a segment whose length is the same as the distance of that point from the origin. ○

**8.K.8.** Consider a chemical compound  $C$  isolated in a container.  $C$  is unstable, with half-time of a molecule equal to  $q$  time units. If there were  $M$  moles of the compound  $C$  in the container at the beginning (i. e., at time  $t = 0$ ), how many moles of it will be there at time  $t \geq 0$ ? ○

**8.K.9.** A 100-gram body lengthens a spring of 5 cm if hung on it. Express the dependency of its position on time  $t$  provided the speed of the body is 10 cm/s when going through the equilibrium point. ○

Further practical problems that lead to differential equations can be found on page 795.

### L. Higher-order differential equations

**8.L.1. Underdamped oscillation.** Now, we will describe a simple model for the movement of a solid object attached to a point with a strong spring. If  $y(t)$  is the deviation of our object from the point  $y_0 = y(0) = 0$ , then we can assume that the acceleration  $y''(t)$  in time  $t$  is proportional to the magnitude of the deviation, yet with the other sign. The proportionality constant  $k$  is called the spring constant. Considering the

### PICARD'S APPROXIMATIONS

**Theorem.** *The unique solution of the equation*

$$y' = f(t, y)$$

*whose right-hand side  $f$  has continuous partial derivatives can be expressed, on a sufficiently small interval, as the limit of step-by-step iterations beginning with the constant function (Picard's approximation):*

$$y_0(t) = y_0, \quad y_{n+1}(t) = L(y_n), \quad n = 1, \dots$$

*It is a uniformly converging sequence of differentiable functions with differentiable limit  $y(t)$ .*

Only the Lipschitz condition is needed for the function  $f$ , so the latter two theorems are true with this weaker assumption as well. It is seen in the next paragraph that continuity of the function  $f$  guarantees the existence of the solution. Yet it is insufficient for the uniqueness.

**8.3.8. Ambiguity of solutions.** We begin with a simple example. Consider the equation

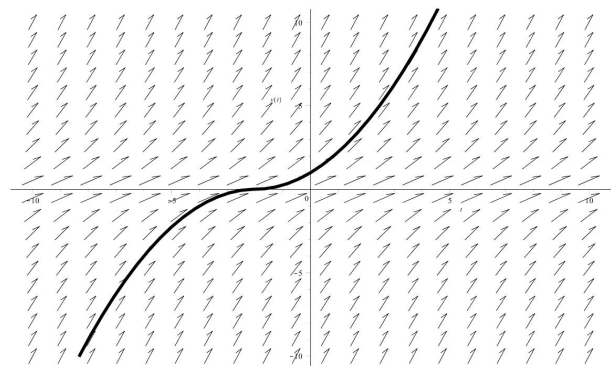
$$y' = \sqrt{|y|}.$$

Separating the variables, the solution is

$$y(t) = \frac{1}{4}(t + C)^2,$$

for positive values  $y$ , with an arbitrary constant  $C$  and  $t + C > 0$ . For the initial values  $(t_0, y_0)$  with  $y_0 \neq 0$ , this is an assignment matching the previous theorem, so there is locally exactly one solution. The solution must apparently remain non-decreasing, hence for negative values  $y_0$ , the solution is the same, only with the opposite sign and  $t + C < 0$ .

However, for the initial condition  $(t_0, y_0) = (t_0, 0)$ , there is not only the already discussed solution continuing to the left of  $t_0$  and to the right, but also the identically zero solution  $y(t) = 0$ . Therefore, these two branches can be glued arbitrarily (see the diagram, where the thick solution can be continued along the  $t$  axis and branch along the parabola at any value  $t$ .)



Nevertheless, the existence of a solution is guaranteed by the following theorem, known as *Peano existence theorem*:

vylepsit obrazek, aby ukazoval vic reseni ...

case  $k = 1$ , we get the so-called oscillation equation

$$y''(t) = -y(t).$$

This equation corresponds to the system of equations

$$x'(t) = -y(t), \quad y'(t) = x(t)$$

from 1. The solution of this system is given by

$$x(t) = R \cos(t - \tau), \quad y(t) = R \sin(t - \tau)$$

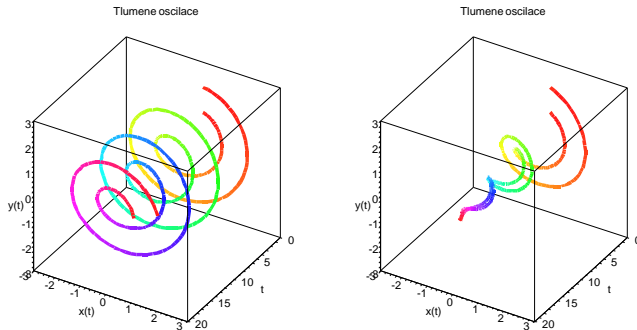
with an arbitrary non-negative constant  $R$ , which determines the maximum amplitude, and a constant  $\tau$ , which determines the initial phase.

Therefore, in order to determine a unique solution, we need to know not only the initial position  $y_0$ , but also the speed of the motion at that moment. These two pieces of information uniquely determine both the amplitude and the initial phase.

Moreover, let us imagine that as a result of the properties of the spring material, there is another force which is directly proportional to the instantaneous speed of our object, with the other sign than the amplitude again. This is expressed by one more term with the first derivative, so our equation is now

$$y''(t) = -y(t) - \alpha y'(t),$$

where  $\alpha$  is a constant which expresses the magnitude of the damping. In the following picture, there are the so-called phase diagrams for solutions with two distinct initial conditions, namely with zero damping on the left, and for the value of the coefficient  $\alpha = 0.3$  on the right.



The oscillations are expressed by the  $y$ -axis values; the  $x$ -axis values describe the speed of the motion.

**8.L.2. Undamped oscillation.** Find the function  $y(t)$  which satisfies the following differential equation and initial conditions:

$$y''(t) + 4y(t) = f(t), \quad y(0) = 0, \quad y'(0) = -1,$$

where the function  $f(t)$  is piecewise continuous:

$$f(t) = \begin{cases} \cos(2t) & \text{for } 0 \leq t < \pi, \\ 0 & \text{for } t \geq \pi. \end{cases}$$

**Solution.** This problem is a model of undamped oscillation of a spring (omitting friction, non-linearities in the toughness of the spring, and other factors) which is initiated by an outer force.

**Theorem.** Consider a function  $f(t, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$  which is continuous on an open set  $U$ . Then for every point  $(t_0, y_0) \in U \supset \mathbb{R}^2$ , there exists a solution of the equation

$$y' = f(t, y)$$

locally in some neighbourhood of  $t_0$ .



**PROOF.** The proof is presented only roughly, with the details left to the reader.

We construct a solution to the right of the initial point  $t_0$ . For this purpose, select a small step  $h > 0$  and label the points

$$t_k = t_0 + kh, \quad k = 1, 2, \dots$$

The value of the derivative  $f(t_0, y_0)$  of the corresponding curve of the solution  $(t, y(t))$  is defined at the initial point  $(t_0, y_0)$ , so a parametrized line with the same derivative can be substituted:

$$y_{(0)}(t) = y_0 + f(t_0, y_0)(t - t_0).$$

Label  $y_1 = y_{(0)}(t_1)$ . Construct inductively the functions and points

$$y_{(k)}(t) = y_k + f(x_k, y_k)(t - t_k), \quad y_{k+1} = y_{(k)}(t_{k+1}).$$

Now, define  $\tilde{y}_h(t)$  by gluing the particular linear parts, i.e., ad picture!!!

$$\tilde{y}_h(t) = y_{(k)}(t) \quad \text{if } t \in [kh, (k+1)h].$$

This is a continuous function, called the *Euler's approximation* of the solution.



It "only" remains to prove that the limit of the functions  $\tilde{y}_h$  for  $h$  approaching zero exists and is a solution. For this, one must observe (as done already in the proof of the theorem on uniqueness and existence of the solution) that  $f(t, y)$  is uniformly continuous on a sufficiently small neighbourhood  $U$  where the solution is sought. For any selected  $\varepsilon > 0$ , a sufficiently small  $\delta$  such that  $|f(t, y) - f(s, z)| < \varepsilon$ , exists whenever  $\|(t - s, y - z)\| < \delta$ .

Especially, all functions  $\tilde{y}_h$  are in the set of uniformly continuous and equicontinuous functions on a sufficiently small interval. By the Arzelà-Ascoli theorem (see paragraph 7.3.15 on page 683), the constructed continuous functions  $\tilde{y}_h$  are all in a compact set of functions. So there exists a sequence of values  $h_n \rightarrow 0$  such that the corresponding sequence of functions  $\tilde{y}_{h_n}$  converges uniformly to a continuous function  $y(n)$ . Write  $\hat{y}_n(t) = \tilde{y}_{h_n}(t)$ , i.e.  $\hat{y}_n \rightarrow y$  uniformly.

For each of the continuous functions  $\tilde{y}_h$ , there are only finitely many points in the interval  $[t_0, t]$  where it is not differentiable, so

$$\hat{y}_n(t) = y_0 + \int_{t_0}^t \hat{y}'_n(s) ds.$$

On the other hand, the derivatives on the particular intervals are constant, so (here,  $k$  is the largest such that  $t_0 + kh_n \leq$



The function  $f(t)$  can be written as a linear combination of Heaviside's function  $u(t)$  and its shift, i. e.,

$$f(t) = \cos(2t)(u(t) - u_\pi(t))$$

Since

$$\mathcal{L}(y'')(s) = s^2 \mathcal{L}(y) - sy(0) - y'(0) = s^2 \mathcal{L}(y) + 1,$$

we get, applying the results of the above exercises 7 and 8 to the Laplace transform of the right-hand side

$$\begin{aligned} s^2 \mathcal{L}(y) + 1 + 4\mathcal{L}(y) &= \mathcal{L}(\cos(2t)(u(t) - u_\pi(t))) \\ &= \mathcal{L}(\cos(2t) \cdot u(t)) - \mathcal{L}(\cos(2t) \cdot u_\pi(t)) \\ &= \mathcal{L}(\cos(2t)) - e^{-\pi s} \mathcal{L}(\cos(2(t + \pi))) \\ &= (1 - e^{-\pi s}) \frac{s}{s^2 + 4}. \end{aligned}$$

Hence,

$$\mathcal{L}(y) = -\frac{1}{s^2 + 4} + (1 - e^{-\pi s}) \frac{s}{(s^2 + 4)^2}.$$

Performing the inverse transform, we obtain the solution in the form

$$y(t) = -\frac{1}{2} \sin(2t) + \frac{1}{4} t \sin(2t) + \mathcal{L}^{-1} \left( e^{-\pi s} \frac{s}{(s^2 + 4)^2} \right).$$

However, by formula (1), we have

$$\begin{aligned} \mathcal{L}^{-1} \left( e^{-\pi s} \frac{s}{(s^2 + 4)^2} \right) &= \frac{1}{4} \mathcal{L}^{-1}(e^{-\pi s} \mathcal{L}(t \sin(2t))) \\ &= (t - \pi) \sin(2(t - \pi)) \cdot H_\pi(t). \end{aligned}$$

Since Heaviside's function is zero for  $t < \pi$  and equal to 1 for  $t > \pi$ , we get the solution in the form

$$y(t) = \begin{cases} -\frac{1}{2} \sin(2t) + \frac{1}{4} t \sin(2t) & \text{for } 0 \leq t < \pi \\ \frac{\pi-2}{4} \sin(2t) & \text{for } t \geq \pi \end{cases}$$

□

**8.L.3.** Find the general solution of the equation

$$y''' - 5y'' - 8y' + 48y = 0.$$

**Solution.** This is a third-order linear differential equation with constant coefficients since it is of the form

$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \dots + a_{n-1} y' + a_n y = f(x)$$

for certain constants  $a_1, \dots, a_n \in \mathbb{R}$ . Moreover, we have  $f(x) \equiv 0$ , i. e., the equation is homogeneous.

First of all, we will find the roots of the so-called characteristic polynomial

$$\lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \dots + a_{n-1} \lambda + a_n.$$

Each real root  $\lambda$  with multiplicity  $k$  corresponds to the  $k$  solutions

$$e^{\lambda x}, x e^{\lambda x}, \dots, x^{k-1} e^{\lambda x}$$

and every pair of complex roots  $\lambda = \alpha \pm i\beta$  with multiplicity  $k$  corresponds to the  $k$  pairs of solutions

$$\begin{aligned} e^{\alpha x} \cos(\beta x), x e^{\alpha x} \cos(\beta x), \dots, x^{k-1} e^{\alpha x} \cos(\beta x), \\ e^{\alpha x} \sin(\beta x), x e^{\alpha x} \sin(\beta x), \dots, x^{k-1} e^{\alpha x} \sin(\beta x). \end{aligned}$$

$t$ , while  $y_j$  and  $t_j$  are the points from the definition of the function  $\hat{y}_{n_n}$ )

$$\hat{y}_n(t) = y_0 + \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} f(t_j, y_j) ds + \int_{t_k}^t f(t_k, y_k) ds.$$

Instead, the equation

$$\hat{y}_n(t) = y_0 + \int_{t_0}^t f(s, \hat{y}_n(s)) ds$$

is wanted, but the difference between this integral and the last two terms in the previous expression is bounded by the possible variation of the function values  $f(t, \hat{y})$  and the lengths of the intervals. By the universal bound for  $f(t, y)$  above, the last integral can be used instead of the actual values in the limit process  $\lim_{n \rightarrow \infty} y_n(t)$ , thereby obtaining

$$\begin{aligned} y(t) &= \lim_{n \rightarrow \infty} \left( y_0 + \int_{t_0}^t f(s, \hat{y}_n(s)) ds \right) \\ &= y_0 + \int_{t_0}^t \left( \lim_{n \rightarrow \infty} f(s, \hat{y}_n(s)) \right) ds \\ &= y_0 + \int_{t_0}^t f(s, y(s)) ds, \end{aligned}$$

where the uniform convergence  $\hat{y}_n(t) \rightarrow y(t)$  is employed.

This proves the theorem. □

**8.3.9. Coupled first-order equations.** The problem of finding the solution of the equation  $y' = f(x, y)$  can



also be viewed as looking for a (parametrized) curve  $(x(t), y(t))$  in the plane where the parametrization of the variable  $x(t) = t$  is fixed beforehand. If this point of view is accepted, then this fixed choice for the variable  $x$  can be forgotten, and the work can be carried out with an arbitrary (finite) number of variables.

In the plane, for instance, such a system can be written in the form

$$x' = f(t, x, y), \quad y' = g(t, x, y)$$

with two functions  $f, g : \mathbb{R}^3 \rightarrow \mathbb{R}$ .

A simple example in the plane might be the system of equations

$$x' = -y, \quad y' = x.$$

It is easily guessed (or at least verified) that there is a solution of this system,

$$x(t) = R \cos t, \quad y(t) = R \sin t,$$

with an arbitrary non-negative constant  $R$ , and the curves of the solution are exactly the parametrized circles with radius  $R$ .

In the general case, the vector notation of the system can be used in the form

$$x' = f(t, x)$$

for a vector function  $x : \mathbb{R} \rightarrow \mathbb{R}^n$  and a mapping  $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ . The validity of the theorem on uniqueness and existence of the solution to such systems can be extended:

Then, the general solution corresponds to all linear combinations of the above solutions.

Therefore, let us consider the polynomial

$$\lambda^3 - 5\lambda^2 - 8\lambda + 48$$

with roots  $\lambda_1 = \lambda_2 = 4$ ,  $\lambda_3 = -3$ . Since we know the roots, we can deduce the general solution as well:

$$y = C_1 e^{4x} + C_2 x e^{4x} + C_3 e^{-3x}, \quad C_1, C_2, C_3 \in \mathbb{R}. \quad \square$$

**8.L.4.** Compute

$$y''' + y'' + 9y' + 9y = e^x + 10 \cos(3x).$$

**Solution.** First, we will solve the corresponding homogeneous equation. The characteristic polynomial is equal to

$$\lambda^3 + \lambda^2 + 9\lambda + 9,$$

with roots  $\lambda_1 = -1$ ,  $\lambda_2 = 3i$ ,  $\lambda_3 = -3i$ . The general solution of the corresponding homogeneous equation is thus

$$y = C_1 e^{-x} + C_2 \cos(3x) + C_3 \sin(3x), \quad C_1, C_2, C_3 \in \mathbb{R}.$$

The solution of the non-homogeneous equation is of the form

$$C_1 e^{-x} + C_2 \cos(3x) + C_3 \sin(3x) + y_p, \quad C_1, C_2, C_3 \in \mathbb{R}$$

for a particular solution  $y_p$  of the non-homogeneous equation.

The right-hand side of the given equation is of a special form. In general, if the non-homogeneous part is given by a function

$$P_n(x) e^{\alpha x},$$

where  $P_n$  is a polynomial of degree  $n$ , then there is a particular solution of the form

$$y_p = x^k R_n(x) e^{\alpha x},$$

where  $k$  is the multiplicity of  $\alpha$  as a root of the characteristic polynomial and  $R_n$  is a polynomial of degree at most  $n$ . More generally, if the non-homogeneous part is of the form

$$e^{\alpha x} [P_m(x) \cos(\beta x) + S_n(x) \sin(\beta x)],$$

where  $P_m$  is a polynomial of degree  $m$  and  $S_n$  is a polynomial of degree  $n$ , there exists a particular solution of the form

$$y_p = x^k e^{\alpha x} [R_l(x) \cos(\beta x) + T_l(x) \sin(\beta x)],$$

where  $k$  is the multiplicity of  $\alpha + i\beta$  as a root of the characteristic polynomial and  $R_l, T_l$  are polynomials of degree at most  $l = \max\{m, n\}$ .

In our problem, the non-homogeneous part is a sum of two functions in the special form (see above). Therefore, we will look for (two) corresponding particular solutions using the method of undetermined coefficients, and then we will add up these solutions. This will give us a particular solution of the original equation (as well as the general solution, then). Let us begin with the function  $y = e^x$ , which has particular solution  $y_{p_1}(x) = Ae^x$  for some  $A \in \mathbb{R}$ . Since

$$y_{p_1}(x) = y'_{p_1}(x) = y''_{p_1}(x) = y'''_{p_1}(x) = Ae^x,$$

substitution into the original equation, whose right-hand side contains only the function  $y = e^x$ , leads to

$$20Ae^x = e^x, \quad \text{i. e.} \quad A = \frac{1}{20}.$$

For the right-hand side with the function  $y = 10 \cos(3x)$ , we are looking for a particular solution in the form

EXISTENCE AND UNIQUENESS FOR SYSTEMS OF ODEs

**Theorem.** Consider functions  $f_i(t, x_1, \dots, x_n) : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, n$ , with continuous partial derivatives. Then, for every point  $(t_0, c) \in \mathbb{R}^{n+1}$ ,  $c = (c_1, \dots, c_n)$ , there exists a maximal interval  $(t_0 - a, t_0 + b)$ , with positive numbers  $a, b \in \mathbb{R}$ , and a unique (vector) function  $x(t) : \mathbb{R} \rightarrow \mathbb{R}^n$  which is the solution of the system of equations

$$x'_1 = f_1(t, x_1, \dots, x_n)$$

⋮

$$x'_n = f_n(t, x_1, \dots, x_n)$$

with the initial condition  $x(t_0) = c$ , i.e.

$$x_1(t_0) = c_1, \dots, x_n(t_0) = c_n.$$

**PROOF.** The proof is almost identical to the one of the existence and uniqueness of the solution for a single equation with a single unknown function as shown in Theorem 8.3.6. The unknown function  $x(t) = (x_1(t), \dots, x_n(t))$  is a curve in  $\mathbb{R}^n$  satisfying the given equation, so its components  $x_i(t)$  are again expressed in terms of integrals

$$x_i(t) = x_i(t_0) + \int_{t_0}^t x'_i(s) ds = c_i + \int_{t_0}^t f_i(s, x(s)) ds.$$

We work with the integral operator  $y \mapsto L(y)$ , this time mapping curves in  $\mathbb{R}^n$  to curves in  $\mathbb{R}^n$ . It is desired to find its fixed point. The proof proceeds in much the same way as in the case 8.3.6. It is only necessary to observe that the size of the vector

$$\|f(t, z_1, \dots, z_n) - f(t, y_1, \dots, y_n)\|$$

is bounded from above by the sum

$$\|f(t, z_1, \dots, z_n) - f(t, y_1, z_2, \dots, z_n)\| + \dots + \|f(t, y_1, \dots, y_{n-1}, z_n) - f(t, y_1, \dots, y_n)\|.$$

It is recommended to go through the proof of Theorem 8.3.6 from this point of view and to think out the details.  $\square$

**8.3.10. Example.** When dealing with models in practice, it is of interest to consider the qualitative behaviour of the solution in dependence on the initial conditions and free parameters of the system

We consider a simple example of a system of first-order equations from this point of view. The standard population model “predator – prey”, was introduced in the 1920s by Lotka and Volterra.

Let  $x(t)$  denote the evolution of the number of individuals in the prey population and  $y(t)$  for the predators. Assume that the increment of the prey would correspond to the Malthusian model (i.e. exponential growth with coefficient  $\alpha$ ) if they were not hunted. On the other hand, assume that the predator would only naturally die out if there were no prey (i.e. exponential decrease with coefficient  $\gamma$ ). Further, consider an

$$y_{p_2}(x) = x [B \cos(3x) + C \sin(3x)].$$

Recall that the number  $\lambda = 3i$  was obtained as a root of the characteristic polynomial. We can easily compute the derivatives

$$\begin{aligned} y'_{p_2}(x) &= [B \cos(3x) + C \sin(3x)] \\ &\quad + x [-3B \sin(3x) + 3C \cos(3x)], \\ y''_{p_2}(x) &= 2[-3B \sin(3x) + 3C \cos(3x)] \\ &\quad + x [-9B \cos(3x) - 9C \sin(3x)], \\ y'''_{p_2}(x) &= 3[-9B \cos(3x) - 9C \sin(3x)] \\ &\quad + x [27B \sin(3x) - 27C \cos(3x)]. \end{aligned}$$

Substituting them into the equation, whose right-hand side contains the function  $y = 10 \cos(3x)$ , we get

$$-18B \cos(3x) - 18C \sin(3x) - 6B \sin(3x) + 6C \cos(3x) = 10 \cos(3x).$$

Confronting the coefficients leads to the system of linear equations

$$-18B + 6C = 10, \quad -18C - 6B = 0$$

with the only solution  $B = -1/2$  and  $C = 1/6$ , i. e.,

$$y_{p_2}(x) = x \left[ -\frac{1}{2} \cos(3x) + \frac{1}{6} \sin(3x) \right].$$

Altogether, the general solution is

$$\begin{aligned} y &= C_1 e^{-x} + C_2 \cos(3x) + C_3 \sin(3x) + \frac{1}{20} e^x \\ &\quad - \frac{1}{2} x \cos(3x) + \frac{1}{6} x \sin(3x), \quad C_1, C_2, C_3 \in \mathbb{R}. \end{aligned}$$

□

**8.L.5.** Determine the general solution of the equation

$$y'' + 3y' + 2y = e^{-2x}.$$

**Solution.** The given equation is a second-order (the highest derivative of the wanted function is of order two) linear (all derivatives are in the first power) differential equation with constant coefficients. First, we solve the homogenized equation

$$y'' + 3y' + 2y = 0.$$

Its characteristic polynomial is

$$x^2 + 3x + 2 = (x + 1)(x + 2),$$

with roots  $x_1 = -1$  and  $x_2 = -2$ . Hence, the general solution of the homogenized equation is

$$c_1 e^{-x} + c_2 e^{-2x},$$

where  $c_1, c_2$  are arbitrary real constants.

Now, using the method of undetermined coefficients, we will find a particular solution of the original non-homogeneous equation. According to the form of the non-homogeneity and since  $-2$  is a root of the characteristic polynomial of the given equation, we are looking for the solution in the form  $y_0 = a x e^{-2x}$  for  $a \in \mathbb{R}$ .

Substituting into the original equation, we obtain

$$a[-4e^{-2x} + 4xe^{-2x} + 3(e^{-2x} - 2xe^{-2x}) + 2xe^{-2x}] = e^{-2x},$$

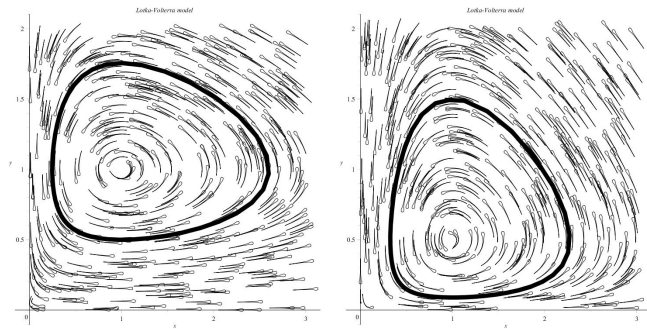
interaction of the predator and the prey which is expected to be proportional to the number of both with a certain coefficient  $\beta$ , which is in the case of the predator, supplemented by a multiplicative coefficient expressing the hunting efficiency.

#### LOTKA-VOLTERRA MODEL

This is a system of two equations,  $x$  models the prey,  $y$  the predator, with positive constants  $\alpha, \beta, \gamma, \delta$

$$\begin{aligned} x' &= \alpha x - \beta y x \\ y' &= -\gamma y + \delta \beta x y. \end{aligned}$$

The diagram illustrates one of typical behaviours of such dynamical systems – the existence of closed orbits on which the system moves in time. These are the thick black ovals, while the “comets” indicate the field at the individual points (i.e. their expected movement). The left illustration corresponds to  $\alpha = 1, \beta = 1, \gamma = 0.3, \delta = 0.3$  and the initial condition  $(x_0, y_0) = (1, 0.5)$  at  $t_0 = 0$  for the solution, while the other illustration comes with  $\alpha = 1, \beta = 2, \gamma = 2, \delta = 1$  and  $(x_0, y_0) = (1, 1.5)$



In both cases, the system is quite stable in the vicinity of the initial condition, and it would be very stable for  $(x_0, y_0) = (1, 1)$  or  $(1, 0.5)$ , respectively. But their development differs in speed — the depicted solution cycles close in the times about  $t = 12$  in the first case and  $t = 5$  in the other one.

It is interesting that the same model captures quite well the development of the unemployment rate in population, considering the employees to be the predators, while the employers play the role of the prey.

Much information about this and other models can be found in the literature.

Next, we are approaching several qualitative results (i.e., featuring properties of the solutions, without knowing them explicitly). They are all easily understandable as statements, but the complexity of their proofs perhaps would be too demanding, at least in the first reading. Thus, the readers are advised to focus on the Theorems (do not ignore them, they are all absolutely essential!) and their rough explanations, but skip the proofs. The exposition will come back to easily proved topics in 8.3.16 on page 785.



hence  $a = -1$ . We have thus found the function  $-xe^{-2x}$  as a particular solution of the given equation. Hence, the general solution is the function space  $c_1e^{-x} + c_2e^{-2x} - xe^{-2x}$ ,  $c_1, c_2 \in \mathbb{R}$ .  $\square$

**8.L.6.** Determine the general solution of the equation

$$y'' + y' = 1.$$

**Solution.** The characteristic polynomial of the given equation is  $x^2 + x$ , with roots 0 and  $-1$ . Therefore, the general solution of the homogenized equation is  $c_1 + c_2e^{-x}$ , where  $c_1, c_2 \in \mathbb{R}$ .

We are looking for a particular solution in the form  $ax$ ,  $a \in \mathbb{R}$  (since zero is a root of the characteristic polynomial). Substituting into the original equation, we get  $a = 1$ . The general solution of the given non-homogeneous equation is  $c_1 + c_2e^{-x} + x$ ,  $c_1, c_2 \in \mathbb{R}$ .  $\square$

**8.L.7.** Determine the general solution of the equation

$$y'' + 5y' + 6y = e^{-2x}.$$

**Solution.** The characteristic polynomial of the equation is  $x^2 + 5x + 6 = (x + 2)(x + 3)$ , its roots are  $-2$  and  $-3$ . The general solution of the homogenized equation is thus  $c_1e^{-2x} + c_2e^{-3x}$ ,  $c_1, c_2 \in \mathbb{R}$ . We are looking for a particular solution in the form  $axe^{-2x}$ , ( $-2$  is a root of the characteristic polynomial),  $a \in \mathbb{R}$ , using the method of undetermined coefficients. Substitution into the original equation yields  $a = 1$ . Hence, the general solution of the given equation is

$$c_1e^{-2x} + c_2e^{-3x} + xe^{-2x}.$$

**8.L.8.** Determine the general solution of the equation

$$y'' - y' = 5.$$

**Solution.** The characteristic polynomial of the equation is  $x^2 - x$ , with roots 1, 0. Therefore, the general solution of the homogenized equation is  $c_1 + c_2e^x$ , where  $c_1, c_2 \in \mathbb{R}$ . We are looking for a particular solution in the form  $ax$ ,  $a \in \mathbb{R}$ , using the method of undetermined coefficients. The result is  $a = -5$ , and the general solution is of the form

$$c_1 + c_2e^x - 5x.$$

**8.L.9.** Solve the equation

$$y'' - 2y' + y = \frac{e^x}{x^2 + 1}.$$

**Solution.** We will solve this non-homogeneous equation using the method of variation of constants. We will thus obtain the solution in the form

$$y = C_1(x)y_1(x) + C_2(x)y_2(x) + \dots + C_n(x)y_n(x),$$

**8.3.11. Stability of systems of equations.** In order to illustrate the stability questions, we discuss just one basic theorem only.

We are interested in the continuity with respect to the  $L_\infty$  norm on the space of functions (i.e. the supremum norm, see 7.3.5). According to the theorem below, the assumption that the partial derivatives of the functions defining the system are continuous (in fact, it suffices to have them Lipschitz), guarantees the continuity of the solutions in dependence on the initial conditions as well as the defining equations themselves. Note however, that as the distance of  $t$  from the initial value  $t_0$  grows, then the error estimates grow exponentially!

Therefore, this result is of a strictly local character. It is not in contradiction with the example of the unstably behaving equation  $y' = ty$  illustrated in paragraph 8.3.3.<sup>8</sup>

Consider two systems of equations written in the vector form

$$(1) \quad x' = f(t, x), \quad y' = g(t, y)$$

and assume that the mappings  $f, g : U \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$  have continuous partial derivatives on an open set  $U$  with compact closure. Such functions must be uniformly continuous and uniformly Lipschitz on  $U$ , so there are the finite values

$$C = \sup_{x \neq y; (t,x), (t,y) \in U} \frac{\|f(t, x) - f(t, y)\|}{\|x - y\|}$$

$$B = \sup_{(t,x) \in U} \|f(t, x) - g(t, x)\|$$

With this notation, the fundamental theorem can be formulated:

**Theorem.** Let  $x(t)$  and  $y(t)$  be two fixed solutions

$$x' = f(t, x(t)), \quad y' = g(t, y(t))$$

of the systems (1) considered above, given by initial conditions  $x(t_0) = x_0$  and  $y(t_0) = y_0$ . Then,

$$\|x(t) - y(t)\| \leq \|x_0 - y_0\| e^{C|t-t_0|} + \frac{B}{C} (e^{C|t-t_0|} - 1).$$

**PROOF.** Without loss of generality,  $t_0 = 0$ . From the expression of the solutions  $x(t)$  and  $y(t)$  as fixed points of the corresponding integral operators follows the estimate

$$\|x(t) - y(t)\| \leq \|x_0 - y_0\| + \int_0^t \|f(s, x(s)) - g(s, y(s))\| ds.$$

The integrand can be further estimated as follows:

$$\begin{aligned} \|f(s, x(s)) - g(s, y(s))\| &\leq \\ &\leq \|f(s, x(s)) - f(s, y(s))\| + \|f(s, y(s)) - g(s, y(s))\| \\ &\leq C \|x(s) - y(s)\| + B \end{aligned}$$

If  $F(t) = \|x(t) - y(t)\|$ ,  $\alpha = \|x_0 - y_0\|$ , then

$$F(t) \leq \alpha + \int_0^t (CF(s) + B) ds.$$

<sup>8</sup>Much more information can be found for example in Gerald Teschl's book *Ordinary Differential Equations and Dynamical Systems*, Graduate Studies in Mathematics, Volume 140, Amer. Math. Soc., Providence, 2012.

where  $y_1, \dots, y_n$  give the general solution of the corresponding homogeneous equation and the functions  $C_1(x), \dots, C_n(x)$  can be obtained from the system

$$\begin{aligned} C_1'(x) y_1(x) + \dots + C_n'(x) y_n(x) &= 0, \\ C_1'(x) y_1'(x) + \dots + C_n'(x) y_n'(x) &= 0, \\ &\vdots \\ C_1'(x) y_1^{(n-2)}(x) + \dots + C_n'(x) y_n^{(n-2)}(x) &= 0, \\ C_1'(x) y_1^{(n-1)}(x) + \dots + C_n'(x) y_n^{(n-1)}(x) &= f(x). \end{aligned}$$

The roots of the characteristic polynomial  $\lambda^2 - 2\lambda + 1$  are  $\lambda_1 = \lambda_2 = 1$ . Therefore, we are looking for the solution in the form

$$C_1(x) e^x + C_2(x) x e^x,$$

considering the system

$$\begin{aligned} C_1'(x) e^x + C_2'(x) x e^x &= 0, \\ C_1'(x) e^x + C_2'(x) [e^x + x e^x] &= \frac{e^x}{x^2 + 1}. \end{aligned}$$

We can compute the unknowns  $C_1(x)$  and  $C_2(x)$  using Cramer's rule. It follows from

$$\begin{aligned} \begin{vmatrix} e^x & x e^x \\ e^x & e^x + x e^x \end{vmatrix} &= e^{2x}, \\ \begin{vmatrix} 0 & x e^x \\ \frac{e^x}{x^2+1} & e^x + x e^x \end{vmatrix} &= -x \frac{e^{2x}}{x^2 + 1}, \\ \begin{vmatrix} e^x & 0 \\ e^x & \frac{e^x}{x^2+1} \end{vmatrix} &= \frac{e^{2x}}{x^2 + 1} \end{aligned}$$

that

$$\begin{aligned} C_1(x) &= - \int \frac{x}{x^2 + 1} dx = -\frac{1}{2} \ln(x^2 + 1) + C_1, & C_1 \in \mathbb{R}, \\ C_2(x) &= \int \frac{dx}{x^2 + 1} = \arctan x + C_2, & C_2 \in \mathbb{R}. \end{aligned}$$

Hence, the general solution is

$$y = C_1 e^x + C_2 x e^x - \frac{1}{2} e^x \ln(x^2 + 1) + x e^x \arctan x, \quad C_1, C_2 \in \mathbb{R}.$$

□

**8.L.10.** Find the only function  $y$  which satisfies the linear differential equation

$$y^{(3)} - 3y' - 2y = 2e^x,$$

with initial conditions  $y(0) = 0, y'(0) = 0, y''(0) = 0$ .

**Solution.** The characteristic polynomial is  $x^3 - 3x - 2$ , with roots 2 and  $-1$  (double). We are looking for a particular solution in the form  $ae^x, a \in \mathbb{R}$ , easily finding out that it is the function  $-\frac{1}{2}e^x$ . The general solution of the given equation is thus

$$c_1 e^{2x} + c_2 e^{-x} + c_3 x e^{-x} - \frac{1}{2} e^x.$$

Such an estimate bound can be exploited further, by the following general result, known as *Gronwall's inequality*. Note the similarity with the general solution of linear equations.

**Lemma.** Assume a real-valued function  $F(t)$  satisfies for all  $t$  in the interval  $[0, t_{max}]$

$$F(t) \leq \alpha(t) + \int_0^t \beta(s) F(s) ds$$

for some real-valued functions  $\alpha(t), \beta(t)$ , with  $\beta(t) \geq 0$ . Then

$$F(t) \leq \alpha(t) + \int_0^t \alpha(s) \beta(s) e^{\int_s^t \beta(r) dr} ds$$

for all  $t \in [0, t_{max}]$ . Moreover, if additionally  $\alpha(t)$  is non-decreasing, then

$$F(t) \leq \alpha(t) e^{\int_0^t \beta(s) ds}.$$

**PROOF OF THE LEMMA.** Write

$$G(t) = e^{-\int_0^t \beta(s) ds}.$$

By the first assumption of the theorem,

$$\begin{aligned} \frac{d}{dt} \left( G(t) \int_0^t \beta(s) F(s) ds \right) &= \\ &= \beta(t) G(t) \left( F(t) - \int_0^t \beta(s) F(s) ds \right) \\ &\leq \alpha(t) \beta(t) G(t). \end{aligned}$$

Integrating with respect to  $t$  and dividing by the non-zero function  $G(t)$  gives

$$\int_0^t \beta(s) F(s) ds \leq \int_0^t \alpha(s) \beta(s) \frac{G(s)}{G(t)} ds,$$

which, having added  $\alpha(t)$  to both sides of the inequality, gives the first proposition of the lemma.

Assuming that  $\alpha(t)$  is non-decreasing, there follows:

$$F(t) \leq \alpha(t) \left( 1 + \int_0^t \beta(s) e^{\int_s^t \beta(r) dr} ds \right).$$

The integrand is a derivative:

$$-\beta(s) e^{\int_s^t \beta(r) dr} = \frac{d}{ds} \left( e^{\int_s^t \beta(r) dr} \right),$$

so

$$\begin{aligned} F(t) &\leq \alpha(t) \left( 1 - \int_0^t \frac{d}{ds} e^{\int_s^t \beta(r) dr} ds \right) \\ &= \alpha(t) \left( 1 + e^{\int_0^t \beta(r) dr} - 1 \right), \end{aligned}$$

and the second proposition of the lemma is also proved. □

Now, the proof of the theorem about continuous dependency on the parameters is easily finished. The bound  $F(t) \leq \alpha + \int_0^t (C F(s) + B) ds$  is already obtained, and using a slightly modified function  $\tilde{F}(t) = F(t) + \frac{B}{C}$ , this yields

$$\tilde{F}(t) \leq \frac{B}{C} + \alpha + \int_0^t C \tilde{F}(s) ds.$$

Substituting into the original conditions, we get the only satisfactory function,

$$\frac{2}{9}e^{2x} + \frac{5}{18}e^{-x} + \frac{1}{3}xe^{-x} - \frac{1}{2}e^x.$$

□

Further problems concerning higher-order differential equations can be found on page 798

**M. Applications of the Laplace transform**

Differential equations with constant coefficients can also be solved using the Laplace transform.

**8.M.1.** Let  $\mathcal{L}(y)(s)$  denote the Laplace transform of a function  $y(t)$ . Integrating by parts, prove that

**Solution.**

$$(1) \quad \begin{aligned} \mathcal{L}(y')(s) &= s\mathcal{L}(y)(s) - y(0) \\ \mathcal{L}(y'')(s) &= s^2\mathcal{L}(y) - sy(0) - y'(0) \end{aligned}$$

and, by induction:

$$\sum_{i=1}^n s^{n-i} y^{(i-1)}(0) . \quad \square$$

**8.M.2.** Find the function  $y(t)$  which satisfies the differential equation

$$y''(t) + 4y(t) = \sin 2t$$

as well as the initial conditions  $y(0) = 0, y'(0) = 0$ .

**Solution.** It follows from the above exercise 7.D.18 that

$$s^2\mathcal{L}(y)(s) + 4\mathcal{L}(y)(s) = \mathcal{L}(\sin 2t)(s).$$

We also have

$$\mathcal{L}(\sin 2t)(s) = \frac{2}{s^2 + 4},$$

i. e.,

$$\mathcal{L}(y)(s) = \frac{2}{(s^2 + 4)^2}.$$

The inverse transform leads to

$$y(t) = \frac{1}{8} \sin 2t - \frac{1}{4}t \cos 2t . \quad \square$$

**8.M.3.** Find the function  $y(t)$  which satisfies the differential equation

$$y''(t) + 6y'(t) + 9y(t) = 50 \sin t$$

and the initial conditions  $y(0) = 1, y'(0) = 4$ .

**Solution.** The Laplace transform yields

$$s^2\mathcal{L}(y)(s) - s - 4 + 6(s\mathcal{L}(y)(s) - 1) + 9\mathcal{L}(y)(s) = 50\mathcal{L}(\sin t)(s),$$

i. e.,

$$(s^2 + 6s + 9)\mathcal{L}(y)(s) = \frac{50}{s^2 + 1} + s + 10,$$

$$\mathcal{L}(y)(s) = \frac{50}{(s^2 + 1)(s + 3)^2} + \frac{s + 10}{(s + 3)^2}.$$

Decomposing the first term to partial fractions, we obtain

$$\frac{50}{(s^2 + 1)(s + 3)^2} = \frac{As + B}{s^2 + 1} + \frac{C}{s + 3} + \frac{D}{(s + 3)^2},$$

This is the assumption of Gronwall's inequality with even constant parameters, so by the second claim of the lemma,

$$F(t) + \frac{B}{C} \leq (\alpha + \frac{B}{C}) e^{\int_0^t C ds},$$

which is the statement

$$F(t) \leq \alpha e^{Ct} + \frac{B}{C}(e^{Ct} - 1)$$

as desired. □

The continuous dependency on both the initial conditions and the potential further parameters in which the function  $f$  would be Lipschitz-continuous follows immediately from the statement of the theorem.

The extremely simple equations in one variable  $x' = ax$ , where  $a$  are small constants, with their exponential solution  $x(t) = e^{at}$  show that better general results cannot be expected.

**8.3.12. Differentiable dependance.** In practical problems, the differentiability of the obtained solutions is often of interest, especially with regard to the initial conditions or other parameters of the system.

In the general vector notation of the system of ordinary equations

$$y' = f(t, y),$$

it can always be supposed that the vector function does not depend explicitly on  $t$ . If it does, then another variable  $y_0$  can be added to the other variables  $y_1, \dots, y_n$ . Then there is the same system of equations for the curve  $\tilde{y}'(t) = (y_0(t), y_1(t), \dots, y_n(t))$  as

$$\begin{aligned} y'_0 &= 1 \\ y'_1 &= f_1(y_0, y_1, \dots, y_n) \\ &\vdots \\ y'_n &= f_n(y_0, y_1, \dots, y_n) \end{aligned}$$

with the initial conditions

$$y_0(t_0) = t_0, y_1(t_0) = x_1, \dots, y_n(t_0) = x_n.$$

Such systems, which do not explicitly depend on time, are called *autonomous systems of ordinary differential equations*.

Without loss of generality, we deal with autonomous systems in finite dimension  $n$ , dependent on parameters  $\lambda$  and with initial conditions

$$(1) \quad y' = f(y, \lambda), y(t_0) = x.$$

Without loss of generality, consider the initial value  $t_0 = 0$ , and write the solution with  $y(0) = x$  in the form  $y(t, x, \lambda)$  to emphasize the dependency on the parameters.

For fixed values of the initial conditions (and the potential parameters  $\lambda$ ), the solution is always once more differentiable than the function  $f$ . This can be derived inductively by applying the chain rule. If  $f$  is continuously differentiable and  $y(t)$  is a solution, then (use the matrix notation where



so

$$50 = (As + B)(s + 3)^2 + C(s^2 + 1)(s + 3) + D(s^2 + 1).$$

Substituting  $s = -3$ , we get

$$50 = 10D \quad \text{hence} \quad D = 5$$

and confronting the coefficients at  $s^3$ , we have

$$0 = A + C, \quad \text{hence} \quad A = -C.$$

Confronting the coefficients at  $s$ , we obtain

$$0 = 9A + 6B + C = 8A + 6B, \quad \text{hence} \quad B = \frac{4}{3}C.$$

Finally, confronting the absolute term, we infer

$$50 = 9B + 3C + D = 12C + 3C + 5$$

$$\text{hence} \quad C = 3, \quad B = 4, \quad A = -3.$$

Since

$$\frac{s + 10}{(s + 3)^2} = \frac{s + 3 + 7}{(s + 3)^2} = \frac{1}{s + 3} + \frac{7}{(s + 3)^2},$$

we have

$$\begin{aligned} \mathcal{L}(y)(s) &= \frac{-3s + 4}{s^2 + 1} + \frac{3}{s + 3} + \frac{5}{(s + 3)^2} + \frac{1}{s + 3} + \frac{7}{(s + 3)^2} \\ &= \frac{-3s}{s^2 + 1} + \frac{4}{s^2 + 1} + \frac{4}{s + 3} + \frac{12}{(s + 3)^2}. \end{aligned}$$

Now, the inverse Laplace transform yields the solution in the form

$$y(t) = -3 \cos t + 4 \sin t + 4e^{-3t} + 12te^{-3t}. \quad \square$$

**8.M.4.** Find the function  $y(t)$  which satisfies the differential equation

$$y''(t) = \cos(\pi t) - y(t), \quad t \in (0, +\infty)$$

and the initial conditions  $y(0) = c_1, y'(0) = c_2$ .

**Solution.** First, we should emphasize that it follows from the theory of ordinary differential equations that this equation has a unique solution. Further, we should recall that

$$\mathcal{L}(f'')(s) = s^2 \mathcal{L}(f)(s) - s \lim_{t \rightarrow 0^+} f(t) - \lim_{t \rightarrow 0^+} f'(t)$$

and

$$\mathcal{L}(\cos(bt))(s) = \frac{s}{s^2 + b^2}, \quad b \in \mathbb{R}.$$

Applying the Laplace transform to the given differential equation then gives

$$s^2 \mathcal{L}(y)(s) - sc_1 - c_2 = \frac{s}{s^2 + \pi^2} - \mathcal{L}(y)(s),$$

i. e.,

$$(1) \quad \mathcal{L}(y)(s) = \frac{s}{(s^2 + 1)(s^2 + \pi^2)} + \frac{c_1 s}{s^2 + 1} + \frac{c_2}{s^2 + 1}.$$

Therefore, it suffices to find a function  $y$  which satisfies (1). Performing partial fraction decomposition, we obtain

$$\frac{s}{(s^2 + 1)(s^2 + \pi^2)} = \frac{1}{\pi^2 - 1} \left( \frac{s}{s^2 + 1} - \frac{s}{s^2 + \pi^2} \right).$$

The above expression of  $\mathcal{L}(\cos(bt))(s)$  and the already proved formula

$$\mathcal{L}(\sin t)(s) = \frac{1}{s^2 + 1}$$

the Jacobi matrix  $D^1 f(y)$  of the mapping  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is multiplied with the column vector  $y'$

$$y'' = D^1 f(y) \cdot y' = D^1 f(y) \cdot f(y)$$

exists and is continuous.

With all the derivatives up to order two continuous, there is an expression for the third derivative:

$$y^{(3)} = D^2 f(y)(f(y), f(y)) + (D^1 f(y))^2 \cdot f(y).$$

Here, the chain rule is used again, starting with the differential of the bilinear mapping of matrix multiplication and viewing the second derivative as a bilinear object evaluated on  $y'$  in both arguments. Think out the argumentation for this and higher orders in detail.

Assume for a while that there is a solution  $y(t, x)$  of the system (1) which is continuously differentiable in the parameters  $x \in \mathbb{R}^n$ , i.e. the initial condition as well, and forget about the further parameters  $\lambda$  for now. Write

$$\Phi(t, x) = D_x^1(y(t, x)),$$

for the Jacobi matrix of all partial derivatives with respect to the coordinates  $x_i$ , which depends on the time  $t$  as well as the initial condition  $x$ . Its derivative  $\Phi'(t, x)$  with respect to  $t$  can be computed using the symmetry of partial derivatives and the chain rule:

$$\begin{aligned} \Phi'(t, x) &= \frac{d}{dt}(D_x^1 y(t, x)) = D_x^1(y'(t, x)) \\ &= D^1 f(y(t, x)) \cdot D_x^1 y(t, x) \\ &= D^1 f(y(t, x)) \cdot \Phi(t, x). \end{aligned}$$

So the derivatives with respect to the initial conditions along the solution  $y(t, x)$  of the system (1) are given as the solutions of a system of  $n^2$  first-order equations with initial condition

$$(2) \quad \Phi'(t, x) = F(t, x) \cdot \Phi(t, x), \quad \Phi(0, x) = E,$$

where  $F(t, x) = D^1 f(y(t, x))$ , and the initial condition comes out from the identity  $y(0, x) = x$ . The unique existence of the solution of this (matrix) system and its continuous dependence on the parameters have already been proved.

The following theorem says that for systems (1) with continuously differentiable right-hand sides  $f$ , the derivatives with respect to the initial condition can be obtained in this way.

#### DIFFERENTIABILITY OF THE SOLUTIONS

**Theorem.** Consider an open subset  $U \subset \mathbb{R}^{n+k}$  and a mapping  $f : U \rightarrow \mathbb{R}^n$  with continuous first derivatives. Then, a system of differential equations dependent on a parameter  $\lambda \in \mathbb{R}^k$  with initial condition at a point  $x \in U$

$$y' = f(y, \lambda), \quad y(0) = x$$

has a unique solution  $y(t, x, \lambda)$ , which is a mapping with continuous first derivatives with respect to each variable.

then yield the wanted solution

$$y(t) = \frac{1}{\pi^2 - 1} (\cos t - \cos(\pi t)) + c_1 \cos t + c_2 \sin t. \quad \square$$

**8.M.5.** Solve the system of differential equations

$$x''(t) + x'(t) = y(t) - y''(t) + e^t, \quad x'(t) + 2x(t) = -y(t) + y'(t) + e^{-t}$$

with the initial conditions  $x(0) = 0, y(0) = 0, x'(0) = 1, y'(0) = 0$ .

**Solution.** Again, we apply the Laplace transform. This, using

$$\mathcal{L}(e^{\pm t})(s) = \frac{1}{s \mp 1},$$

transforms the first equation to

$$s^2 \mathcal{L}(x)(s) - s \lim_{t \rightarrow 0+} x(t) - \lim_{t \rightarrow 0+} x'(t) + s \mathcal{L}(x)(s) - \lim_{t \rightarrow 0+} x(t) =$$

$$\mathcal{L}(y)(s) - \left( s^2 \mathcal{L}(y)(s) - s \lim_{t \rightarrow 0+} y(t) - \lim_{t \rightarrow 0+} y'(t) \right) + \frac{1}{s-1}$$

and the second one to

$$s \mathcal{L}(x)(s) - \lim_{t \rightarrow 0+} x(t) + 2 \mathcal{L}(x)(s) = -\mathcal{L}(y)(s) + s \mathcal{L}(y)(s) - \lim_{t \rightarrow 0+} y(t) + \frac{1}{s+1}.$$

Evaluating the limits (according to the initial conditions), we obtain the linear equations

$$s^2 \mathcal{L}(x)(s) - 1 + s \mathcal{L}(x)(s) = \mathcal{L}(y)(s) - s^2 \mathcal{L}(y)(s) + \frac{1}{s-1}$$

and

$$s \mathcal{L}(x)(s) + 2 \mathcal{L}(x)(s) = -\mathcal{L}(y)(s) + s \mathcal{L}(y)(s) + \frac{1}{s+1}$$

with the only solution

$$\mathcal{L}(x)(s) = \frac{2s-1}{2(s-1)(s+1)^2}, \quad \mathcal{L}(y)(s) = \frac{3s}{2(s^2-1)^2}.$$

Once again, we perform partial fraction decomposition, getting

$$\mathcal{L}(x)(s) = \frac{1}{8} \frac{1}{s-1} + \frac{3}{4} \frac{1}{(s+1)^2} - \frac{1}{8} \frac{1}{s+1} = \frac{3}{4} \frac{1}{(s+1)^2} + \frac{1}{4} \frac{1}{s^2-1}.$$

Since we have already computed that

$$\mathcal{L}(t e^{-t})(s) = \frac{1}{(s+1)^2}, \quad \mathcal{L}(\sinh t)(s) = \frac{1}{s^2-1},$$

$$\mathcal{L}(t \sinh t)(s) = \frac{2s}{(s^2-1)^2},$$

we get

$$x(t) = \frac{3}{4} t e^{-t} + \frac{1}{4} \sinh t, \quad y(t) = \frac{3}{4} t \sinh t.$$

We definitely advise the reader to verify that these functions of  $x$  and  $y$  are indeed the wanted solution. The reason is that the Laplace transforms of the functions  $y = e^t, y = \sinh t$  and  $y = t \sinh t$  were obtained only for  $s > 1$ .  $\square$

**8.M.6.** Find the solution of the following system of differential equations:

$$x'(t) = -2x(t) + 3y(t) + 3t^2,$$

$$y'(t) = -4x(t) + 5y(t) + e^t, \quad x(0) = 1, y(0) = -1$$

**Solution.**

$$\mathcal{L}(x')(s) = \mathcal{L}(-2x + 3y + 3t^2)(s),$$

$$\mathcal{L}(y')(s) = \mathcal{L}(-4x + 5y + e^t)(s).$$



**PROOF.** Consider a general system dependent on parameters, but viewed as an ordinary autonomous system with no parameters. More explicitly, consider the parameters to be additional space variables and add (vector) conditions  $\lambda'(t) = 0$  and  $\lambda(0) = \lambda$ . Therefore, the theorem is proved for autonomous systems with no further parameters. There is dependency on the initial conditions.

Just as in the proof of the fundamental existence theorem 8.3.6, build on the expression of the solutions as fixed points of the integral operators and prove that the expected derivative, as discussed above, enjoys the properties of the differential. Fix a point  $x_0$  as the initial condition, together with a small neighbourhood  $x_0 \in V$ , which if necessary can be further decreased during the following estimates, so that

$$\|f(y) - f(z)\| \leq C \|y - z\|$$

on this neighbourhood by the Lipschitz property. It is already deduced that if the derivative

$$\Phi(t, x) = D_x^1 y(t, x)$$

of the solution  $y(t, x)$  exists, then it must be uniquely given by the equation (2) with the proper initial conditions. Therefore, define  $\Phi(t, x)$  by this equation and examine the expression

$$G(t, h) = \|y(t, x_0 + h) - y(t, x_0) - \Phi(t, x_0)(h)\|$$

with small increments  $h \in \mathbb{R}^n$ . In order to prove that the continuous derivative exists, it is necessary to show that

$$\lim_{h \rightarrow 0} \frac{1}{\|h\|} G(t, h) = 0.$$

Several estimates are needed for this purpose.

First, from the latter theorem 8.3.11 about continuous dependence on initial conditions, the estimate

$$\|y(t, x_0 + h) - y(t, x_0)\| \leq \|h\| e^{C|t|}$$

follows immediately. In the next step, use Taylor's expansion of  $f$  with remainder:

$$f(y) - f(z) = D^1 f(z) \cdot (y - z) + R(y, z),$$

where  $R(y, z)$  satisfies

$$\frac{R(y, z)}{\|y - z\|} \rightarrow 0 \text{ for } \|y - z\| \rightarrow 0.$$

This implies the crucial estimate. In the first equality substitute in the expression of solutions in terms of fixed points of the integral operators. Next, exploit the definition of the mapping  $\Phi(t, x_0)$  in terms of its derivative (write  $F(t, x) = D^1 f(y(t, x))$  again and notice that its initial condition  $\Phi(0, x)(h) = h$  implies the vanishing of the  $h$



The left-hand sides can be written using (1), while the right-hand sides can be rewritten thanks to linearity of the  $\mathcal{L}$  operator. Since  $\mathcal{L}(3t^2)(s) = \frac{6}{s^3}$  and  $\mathcal{L}(e^t)(s) = \frac{1}{s-1}$ , we get the system of linear equations

$$\begin{aligned} s\mathcal{L}(x)(s) - 1 &= -2\mathcal{L}(x)(s) + 3\mathcal{L}(y)(s) + \frac{6}{s^3}, \\ s\mathcal{L}(y)(s) + 1 &= -4\mathcal{L}(x)(s) + 5\mathcal{L}(y)(s) + \frac{1}{s-1}. \end{aligned}$$

In matrices, this is  $\mathbf{A}(s)\hat{\mathbf{x}}(s) = \mathbf{b}(s)$ , where

$$\mathbf{A}(s) = \begin{pmatrix} s+2 & -3 \\ 4 & s-5 \end{pmatrix}, \quad \hat{\mathbf{x}}(s) = \begin{pmatrix} \mathcal{L}(x)(s) \\ \mathcal{L}(y)(s) \end{pmatrix} \quad \text{and} \quad \mathbf{b}(s) = \begin{pmatrix} 1 + \frac{6}{s^3} \\ -1 + \frac{1}{s-1} \end{pmatrix}.$$

Cramer's rule says that

$$\mathcal{L}(x)(s) = \frac{|\mathbf{A}_1|}{|\mathbf{A}|}, \quad \mathcal{L}(y)(s) = \frac{|\mathbf{A}_2|}{|\mathbf{A}|}, \quad \text{where}$$

$$|\mathbf{A}| = \begin{vmatrix} s+2 & -3 \\ 4 & s-5 \end{vmatrix} = s^2 - 3s + 2,$$

$$|\mathbf{A}_1| = \begin{vmatrix} 1 + \frac{6}{s^3} & -3 \\ -1 + \frac{1}{s-1} & s-5 \end{vmatrix} = (s-5)\left(1 + \frac{6}{s^3}\right) + 3\left(-1 + \frac{1}{s-1}\right)$$

$$|\mathbf{A}_2| = \begin{vmatrix} s+2 & 1 + \frac{6}{s^3} \\ 4 & -1 + \frac{1}{s-1} \end{vmatrix} = (s+2)\left(-1 + \frac{1}{s-1}\right) - 4 - \frac{24}{s^3}.$$

Hence,

$$\mathcal{L}(x)(s) = \frac{1}{(s-1)(s-2)} \left( \frac{(s-5)(s^3+6)}{s^3} - 3\frac{s-2}{s-1} \right),$$

$$\mathcal{L}(y)(s) = \frac{1}{(s-1)(s-2)} \left( \frac{(s+2)(2-s)}{s-1} - \frac{4s^3+24}{s^3} \right).$$

Decomposing to partial fractions, the Laplace images of the solutions can be expressed as follows:

$$\mathcal{L}(x)(s) = -\frac{39}{2s^2} - \frac{3}{(s-1)^2} + \frac{28}{s-1} - \frac{21}{4(s-2)} - \frac{15}{s^3} - \frac{87}{4s},$$

$$\mathcal{L}(x)(s) = -\frac{18}{s^2} - \frac{3}{(s-1)^2} + \frac{27}{s-1} - \frac{7}{s-2} - \frac{12}{s^3} - \frac{21}{s}.$$

Now, the inverse transform yields the solution of this Cauchy problem:

$$\begin{aligned} x(t) &= -\frac{39}{2}t - 3te^t + 28e^t - \frac{21}{4}e^{2t} - \frac{15}{2}t^2 - \frac{87}{4}, \\ y(t) &= -18t - 3te^t + 27e^t - 7e^{2t} - 6t^2 - 21. \quad \square \end{aligned}$$

## N. Numerical solution of differential equations

Now, we present two simple exercises on applying the Euler method for solving differential equations.

**8.N.1.** Use the Euler method to solve the equation  $y' = -y^2$  with the initial condition  $y(1) = 1$ . Determine the approximate solution on the interval  $[1, 3]$ . Try to estimate for which value  $h$  of the step is the error less than one tenth.

**Solution.** The Euler method for the considered equation is given by

$$y_{k+1} = y_k - h \cdot y_k^2$$

for

$$x_0 = 1, \quad y_0 = 1, \quad x_k = x_0 + k \cdot h, \quad y_k = y(x_k).$$

We begin the procedure with step value  $h = 1$  and halve it in each iteration. The estimate for the "sufficiency" of  $h$  will be made somewhat imprecisely by comparing two adjacent

summand).

$$\begin{aligned} G(t, h) &= \left\| x_0 + h + \int_0^t f(y(s, x_0 + h)) ds - x_0 \right. \\ &\quad \left. - \int_0^t f(y(s, x_0)) ds - \Phi(t, x_0)(h) \right\| \\ &= \left\| h + \int_0^t \left( f(y(s, x_0 + h)) - f(y(s, x_0)) \right) \right. \\ &\quad \left. - F(s, x_0)\Phi(s, x_0)(h) ds - h \right\| \\ &\leq \int_0^t \|f(y(s, x_0 + h)) - f(y(s, x_0)) \\ &\quad - F(s, x_0)\Phi(s, x_0)(h)\| ds \\ &\leq \int_0^t \|F(s, x_0)\| \|y(s, x_0 + h) - y(s, x_0) - \Phi(s, x_0)(h)\| ds \\ &\quad + \int_0^t \|R(y(s, x_0 + h), y(s, x_0))\| ds, \end{aligned}$$

where the norm on the matrices is taken as the maximum of the absolute values of their entries.

Since  $F(t, x)$  is continuous, there is a uniform bound of its norm in the neighbourhood  $V$  given by

$$\|F(t, x_0)\| \leq B,$$

for all  $|t| < T$  with a sufficiently small  $T$  to ensure the solutions remain in the neighbourhood  $V$ . At the same time, for any fixed constant  $\varepsilon > 0$ , there is a bound  $\|h\| < \delta$  for which the remainder  $R$  satisfies

$$\begin{aligned} \|R(y(t, x_0 + h), y(t, x_0))\| &\leq \varepsilon \|y(t, x_0 + h) - y(t, x_0)\| \\ &\leq \|h\| \varepsilon e^{CT}. \end{aligned}$$

Therefore, the estimate on  $G(t, h)$  can be improved as follows:

$$G(t, h) \leq B \int_0^t G(s, h) ds + \varepsilon \|h\| e^{CT}.$$

Gronwall's lemma now gives

$$G(t, h) \leq \varepsilon \|h\| e^{(C+B)T}.$$

This implies that  $\lim_{h \rightarrow 0} \frac{1}{\|h\|} G(t, h) = 0$  as requested.  $\square$

In the same way, it can be proved that continuous differentiability of the right-hand side up to order  $k$  (inclusive) guarantees the same order of differentiability of solutions in all input parameters.

**8.3.13. The analytic case.** Let us pay additional attention to the case when the right hand side  $f$  of the system of equations

$$(1) \quad y' = f(y), \quad y(t_0) = y_0$$

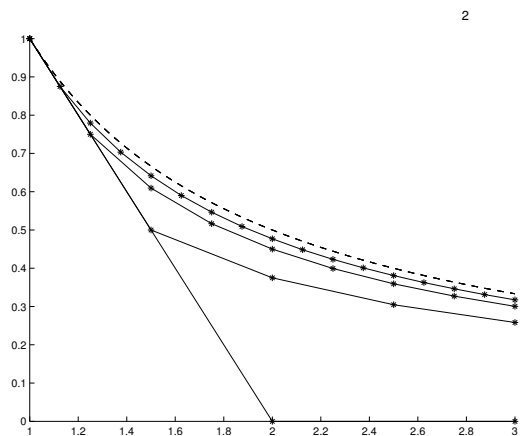
is analytic in all arguments (i.e., a convergent multidimensional power series  $f(y) = \sum_{|\alpha|=0}^{\infty} \frac{1}{\alpha!} \frac{\partial f^{\alpha}}{\partial y^{\alpha}} y^{\alpha}$ , see 8.1.15). Exactly as in the previous discussion, we may hide the time variable  $t$  as well as further parameters in the variables.

approximate values of the function  $y$  at common points, terminating the procedure if the maximum of the absolute difference of these values is not greater than the tolerated error (0.1).

The results

$h_0 = 1$					
$y^{(0)} = (1 \ 0 \ 0)$					
$h_1 = 0.5$					
$y^{(1)} = (1 \ 0.5 \ 0.375 \ 0.3047 \ 0.2583)$					
Maximal difference: 0.375.					
$h_2 = 0.25$					
$y^{(2)} = (1.0000 \ 0.7500 \ 0.6094 \ 0.5165 \ 0.4498$					
$0.3992 \ 0.3594 \ 0.3271 \ 0.3004)$					
Maximal difference: 0.1094.					
$h_3 = 0.125$					
$y^{(3)} = (1.0000 \ 0.8750 \ 0.7793 \ 0.7034 \ 0.6415$					
$0.5901 \ 0.5466 \ 0.5092 \ 0.4768 \ 0.4484$					
$0.4233 \ 0.4009 \ 0.3808 \ 0.3627$					
$0.3462 \ 0.3312 \ 0.3175)$					
Maximal difference: 0.0322.					

Using suitable software, the following graphical representation of the results can be obtained, where the dashed curve corresponds to the exact solution, which is the function  $y = 1/x$ .



2

□

**8.N.2.** Using the Euler method, solve the equation  $y' = -2y$  with the initial condition  $y(0) = 1$  and step value  $h = 1$ . Explain the phenomenon which occurs here and suggest another procedure.

**Solution.** In this case, the Euler method is given by

$$y_{k+1} = y_k - h \cdot 2y_k = -y_k.$$

For the initial condition  $y_0 = 1$ , we get the alternating values  $\pm 1$  as the result. This is a typical manifestation of the instability of this method for large step values  $h$ . If the step cannot be reduced for some reasons (for instance, when processing digital data, the step value is fixed), better results can be achieved by the so-called *implicit Euler method*. For a general equation  $y' = f(x, y)$ , that is given by the formula

$$y_{k+1} = y_k + h \cdot f(x_{k+1}, y_{k+1}).$$

The famous theorem below says that the solution of the most general system with analytic right-hand side is analytic in all the parameters as well (including the initial conditions).

ODE VERSION OF CAUCHY-KOVALEVSKAYA THEOREM

**Theorem.** Assume  $f(y)$  is a real analytic vector valued function on a domain in  $\mathbb{R}^n$  and consider the differential equation (1). Then the unique solution of this initial problem is real analytic, including the dependency on the initial condition.

**PROOF.** The idea of the proof is identical as in the simple one-dimensional case in 6.2.22. As we saw in the beginning of the previous paragraph, there are universal (multidimensional) polynomial expressions for all derivatives of the vector function  $y(t)$  in terms of the partial derivatives of the vector function  $f$ . If we expand them in terms of the individual partial derivatives of the mapping  $f$  all of their coefficients are obviously non-negative. Let us write again

$$y^{(k)}(0) = P_k(f(y(0)), \dots, \partial_\beta f(y(0)), \dots)$$

for these multivariate vector valued polynomials (the multi-indices  $\beta$  in the arguments are all of size up to  $k - 1$ ).

Without loss of generality we may consider the initial condition  $t_0 = 0, y(0) = 0$ . Indeed, constant shifts of the variables (say  $z = y - y_0, x = t - t_0$ ) transform the general case to this one. Once we know that the components of the solution are power series, the transformed quantities will be analytic too, including the dependency on the values of the incital conditions.

In order to prove that the solution to the problem  $y' = f(y), y(0) = 0$  is analytic on a neighborhood of the origin, we shall again look for a majorant  $g$  for the vector equation  $y' = f(y)$ , i.e. we want an analytic function on a neighborhood of the origin  $0 \in \mathbb{R}^n$  with  $\partial_\alpha g(0) \geq |\partial_\alpha f(0)|$ , for all multi-indices  $\alpha$ . Then, by the universal computations of all the coefficients of the power series  $y(t) = \sum_{k=0}^\infty \frac{1}{k!} y^{(k)}(0) t^k$  solving potentially our problem, and similarly for  $z' = g(z)$ , the convergence of the series for  $z$  implies the same for  $y$ :

$$\begin{aligned} z^{(k)}(0) &= P_k(g(0), \dots, \partial_\beta g(0), \dots) \\ &\geq P_k(|f(0)|, \dots, |\partial_\beta f(0)|) \geq |y^{(k)}(0)|. \end{aligned}$$

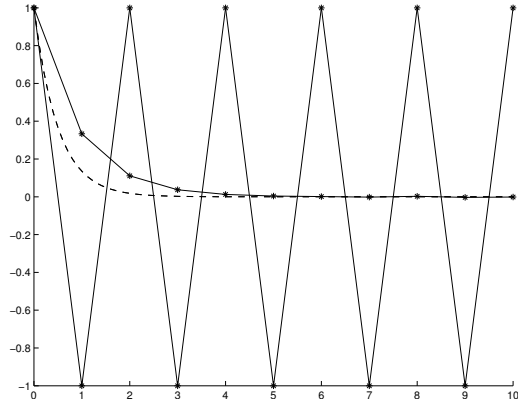
As usual, knowing already how to find a majorant in a simpler case, we try to apply a straightforward modification.

By the analycity of  $f$ , for  $r > 0$  small enough there is a constant  $C$  such that  $|\frac{1}{\alpha!} \partial_\alpha f_i(0) r^{|\alpha|}| \leq C$ , for all  $i = 1, \dots, n$  and mutli-indices  $\alpha$ . This means  $|\partial_\alpha f_i(0)| \leq C \frac{\alpha!}{r^{|\alpha|}}$ . In the 1-dimensional case, we considered the multiple of a geometric series  $g(z) = C \frac{r}{r-z}$  with the right derivatives  $g^{(n)} = C \frac{n!}{r^n}$ . Now the most similar mapping is  $g(z_1, \dots, z_n) = (g_1(z_1, \dots, z_n), \dots, g_n(z_1, \dots, z_n))$  with

In general, we thus have to solve a non-linear equation in each step. However, in our problem, we get

$$y_{k+1} = y_k - 2h \cdot y_{k+1},$$

so we have  $y_{k+1} = \frac{1}{3}y_k$  for  $h = 1$ . Again, the obtained results can be represented graphically, including the exact solution of the equation.



all the components  $g_i$  equal to

$$h : \mathbb{R}^n \rightarrow \mathbb{R}, \quad h(z_1, \dots, z_n) = C \frac{r}{r - z_1 - \dots - z_n}.$$

Then the values of all the partial derivatives with  $|\alpha| = k$  at  $z = 0$  are

$$\partial_\alpha h(0) = Crk!(r - z_1 - \dots - z_n)^{-k-1}|_{z=0} = C \frac{k!}{r^k},$$

exactly as suitable. (Check the latter simple computation yourself!)

So it remains to prove that the majorant system  $z' = g(z)$  has got the converging power series solution  $z$ . Obviously, by the symmetry of  $g$  (all components equal to the same  $h$  and  $h$  is symmetric in the variables  $z_i$ ), also the solution  $z$  with  $z(0) = 0$  must have all the components equal (the system does not see any permutation of the variables  $z_i$  at all). Let us write  $z_i(t) = u(t)$  for the common solution components. With this ansatz,

$$u'(t) = h(u(t), u(t), \dots, u(t)) = C \frac{r}{r - nu(t)}.$$

This is nearly exactly the same equation as the one in 6.2.22 and we can easily see its solution with  $u(0) = 0$ :

□

$$u = \frac{r}{n} \left( 1 - \sqrt{1 - \frac{2nCt}{r}} \right).$$

Clearly, this is an analytic solution and the proof is finished. □

**8.3.14. Vector fields and their flows.** Before going to



higher-order equations, pause to consider systems of first-order equations from the geometrical point of view. When drawing illustrations of solutions earlier, we already viewed the right hand side of an autonomous system as a “field of vectors”  $f(x) \in \mathbb{R}^n$ . This shows how fast and in which direction the solution should move in time.

This can be formalized. A *tangent vector* with a footpoint  $x \in \mathbb{R}^n$  is a couple  $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$ . The set of all vectors with footpoints in an open set  $U \subset \mathbb{R}^n$  is called the *tangent bundle*  $TU$ , with the footpoint projection  $p : (x, v) \mapsto x$ . A *vector field*  $X$  defined on an open set  $U \subset \mathbb{R}^n$  is a mapping  $X : U \rightarrow TU$  which is a section of the projection  $p$ , i.e.,  $p \circ X = \text{id}_U$ . The *derivative in the direction of the vector field*  $X$  is defined for all differentiable functions  $g$  on  $U$  by

$$X(g) : U \rightarrow \mathbb{R}, \quad X(g)(x) = d_{X(x)}g = dg(x)(X(x)).$$

So the vector field  $X$  is a first order linear differential operator mapping functions into functions. Apply pointwise the properties of directional derivative to obtain the *derivative rule* (also called the *Leibniz rule*) for products of functions:

$$(1) \quad X(gh) = hX(g) + gX(h).$$

In fixed coordinates,  $X(x) = (X_1(x), \dots, X_n(x))$  and

$$X(g)(x) = X_1(x) \frac{\partial g}{\partial x_1}(x) + \dots + X_n(x) \frac{\partial g}{\partial x_n}(x).$$

Clearly, there are the special vector fields with coordinate functions equal to zero except for one function  $X_i$  which is identically one. Such a field then corresponds to the partial derivatives with respect to the variable  $x_i$ . This is also matched by the common notation  $\frac{\partial}{\partial x_i}$  for such vector fields and in general,

$$X(x) = X_1(x) \frac{\partial}{\partial x_1} + \cdots + X_n(x) \frac{\partial}{\partial x_n}.$$

**Remark.** Actually, each derivative on functions, i.e., a linear operator  $D$  satisfying (1), is given by a unique vector field. This may be seen as follows. First  $D(1) = D(1 \cdot 1) = 2D(1)$  and, thus,  $D(c) = 0$  for constant functions. Next, each function  $f(x)$  can be written on a neighborhood of a point  $q \in \mathbb{R}^n$  as  $f(x) = f(q) + \int_0^1 \frac{d}{dt} f(q + t(x - q)) dt = f(q) + \sum_{i=1}^n \int_0^1 \frac{\partial f}{\partial x_i}(q + t(x - q)) dt (x_i - q_i) = f(q) + \sum_{i=1}^n \alpha_i(x)(x_i - q_i)$ . Thus,  $D(f) = 0 + \sum_{i=1}^n D(\alpha_i)(q_i - q_i) + \sum_{i=1}^n \alpha_i(q) D(x_i) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(q) D(x_i)$ . Defining the components  $X_i = D(x_i)$  of the vector field  $X$ , we have obtained  $D$  as the derivative in the direction of  $X$ .

We shall write  $\mathcal{X}(U)$  for the set of all smooth vector fields on  $U$ , i.e. those with all components  $X_i$  smooth. Vector fields  $\frac{\partial}{\partial x_i}$  can be perceived as generators of  $\mathcal{X}(U)$ , admitting smooth functions as the coefficients in linear combinations.

We return to the problem of finding the solution of a system of equations. Rephrase it equivalently as finding a curve which satisfies

$$x'(t) = X(x(t))$$

for each value  $x(t)$  in the domain of the vector field  $X$ . In words: the tangent vector of the curve is given, at each of its points, by the vector field  $X$ . Such a curve is called an *integral curve* of the vector field  $X$ , and the mapping

$$\text{Fl}_t^X : \mathbb{R}^n \rightarrow \mathbb{R}^n,$$

defined at a point  $x_0$  as the value of the integral curve  $x(t)$ , satisfying  $x(0) = x_0$  is called the *flow of the vector field*  $X$ . The theorem about existence and uniqueness of the solution of the systems of equations says (cf. 8.3.6) that for every continuously differentiable vector field  $X$ , its flow exists at every point  $x_0$  of the domain for sufficiently small values of  $t$ . The uniqueness guarantees that

$$\text{Fl}_{t+s}^X(x) = \text{Fl}_t^X \circ \text{Fl}_s^X(x),$$

whenever both sides exist. In particular, the mappings  $\text{Fl}_s^X$  and  $\text{Fl}_t^X$  always commute.

Moreover, the mapping  $\text{Fl}_t^X(x)$  with a fixed parameter  $t$  is differentiable at all points  $x$  where it is defined, cf. 8.3.12.

If a vector field  $X$  is defined on all of  $\mathbb{R}^n$ , and if its support is compact (i.e.,  $X(x) = 0$  off a compact set  $K \subset \mathbb{R}^n$ ), then its flow clearly exists at all points and for all values of  $t$ . Vector fields with flows existing for all  $t \in \mathbb{R}$  are called *complete*. The flow of a complete vector field consists of (mutually commuting) diffeomorphisms  $\text{Fl}_t^X : \mathbb{R}^n \rightarrow \mathbb{R}^n$  with inverse diffeomorphisms  $\text{Fl}_{-t}^X$ .

A simple example of a complete vector field is the field  $X(x) = \frac{\partial}{\partial x_1}$ . Its flow is given by

$$\text{Fl}_t^X(x_1, \dots, x_n) = (x_1 + t, x_2, \dots, x_n).$$

On the other hand, the vector field  $X(t) = t^2 \frac{d}{dt}$  on the one-dimensional space  $\mathbb{R}$  is not complete as the solutions  $x(t)$  of the corresponding equation  $dx = x^2 dt$  are of the form  $x(t) = \frac{1}{C-t}$ , except for the initial condition  $x(0) = 0$ , so they “run away” towards infinite values in a finite time.

The points  $x_0$  in the domain of a vector field  $X : U \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$  where  $X(x_0) = 0$  are called *singular points* of the vector field  $X$ . Clearly  $\text{Fl}_t^X(x_0) = x_0$  for all  $t$  at all singular points.

**8.3.15. Local qualitative description.** The description of vector fields as assigning the tangent vector in the modelling space to each point of the Euclidean space is independent of the coordinates. It follows that the flows exhibit a geometric concept which must be coordinate-free.

It is necessary to know what happens to the fields and their flows, when coordinates are transformed. Suppose  $y = F(x)$  is such a transformation with  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  (or on some smaller domain there). Then the solutions  $x(t)$  to a system  $x' = X(x)$  satisfy  $x'(t) = X(x(t))$ , and in the transformed coordinates this reads

$$\begin{aligned} y'(t) &= (F(x(t)))'(t) = D^1 F(x(t)) \cdot x'(t) \\ &= D^1 F(x(t)) \cdot X(x(t)). \end{aligned}$$

This means that the “transformed field”  $Y$  in the new coordinates is  $Y(F(x)) = D^1 F(x) \cdot X(x)$ . At the same time, the flows of these vector fields are related as follows:

$$\text{Fl}_t^Y \circ F(x) = F \circ \text{Fl}_t^X(x).$$

By fixing  $x = x_0$  and writing  $x(t) = \text{Fl}_t^X(x_0)$ , the curve  $F(x(t))$  is the unique solution for the system of equations  $y' = Y(y)$  with initial condition  $y_0 = F(x_0)$ , which equals the right-hand side.

The following theorem offers a geometric local qualitative description of all solutions of systems of first order ordinary differential equations in a neighbourhood of each point  $x$  which is not singular.

THE FLOWBOX THEOREM

**Theorem.** *If  $X$  is a differentiable vector field defined on a neighbourhood of a point  $x_0 \in \mathbb{R}^n$  and  $X(x_0) \neq 0$ , then there exists a transformation of coordinates  $F$  such that in the new coordinates  $y = F(x)$ , the vector field  $X$  is given as the field  $\frac{\partial}{\partial y_1}$ .*

**PROOF.** Construct a diffeomorphism  $F$  with the required properties, step by step. Geometrically, the essence of the proof can be summarized as follows: first select a hypersurface which goes through the point  $x_0$  and is complementary to the directions  $X(x)$  near to  $x_0$ . Then fix the coordinates

on it, and finally, extend them to some neighbourhood of the point  $x_0$  using the flow of the field  $X$ .

Without loss of generality, move the point  $x_0$  to the origin by a translation. Then by a suitable linear transformation on  $\mathbb{R}^n$ , set  $X(0) = \frac{\partial}{\partial x_1}(0)$ . Let us write  $\xi$  for the vector field in question, independent of any coordinates.



In the standard coordinates on  $\mathbb{R}^n$ , write the flow  $\text{Fl}_t^\xi$  of the field  $X$  going through the point  $(x_1, \dots, x_n)$  at time  $t = 0$  as  $x_i(t) = \varphi_i(t, x_1, \dots, x_n)$ ,  $i = 1, \dots, n$ . Next, define the new coordinates  $y = (y_1, \dots, y_n)$  by  $y = \text{Fl}_{y_1}^\xi(0, y_2, \dots, y_n)$ , which corresponds to the inverse transformation  $F^{-1}$  to the diffeomorphism  $F = (f_1, \dots, f_n)$  with the components

$$f_i(x_1, \dots, x_n) = \varphi_i(x_1, 0, x_2, \dots, x_n).$$

This follows exactly the strategy using the hypersurface  $x_1 = 0$ . Since  $\xi(0, \dots, 0) = \frac{\partial}{\partial x_1}$

$$\begin{aligned} \frac{\partial F}{\partial x_1}(0, \dots, 0) &= \frac{d}{dt}\bigg|_0 (\varphi_1(t, 0, \dots, 0), \dots, \varphi_n(t, 0, \dots, 0)) \\ &= (1, 0, \dots, 0), \end{aligned}$$

while the flow  $\text{Fl}_0^\xi$  at the time  $t = 0$  yields

$$\varphi_i(0, 0, x_2, \dots, x_n) = (0, x_2, \dots, x_n),$$

and in particular

$$\frac{\partial F}{\partial x_i}(0, \dots, 0) = (0, \dots, 1, \dots, 0), \quad i = 2, \dots, n.$$

Therefore, the Jacobi matrix of the mapping  $F$  at the origin is the identity matrix  $E$ , so it is a transformation of coordinates on some neighbourhood (see the inverse mapping theorem in paragraph 8.1.23).

Directly from the definition of the mapping  $F^{-1}$  we can compute

$$\begin{aligned} F^{-1}(\text{Fl}_t^\xi(y)) &= F^{-1}(\text{Fl}_{t+y_1}^\xi(0, y_2, \dots, y_n)) \\ &= F^{-1} \circ F(t + y_1, \dots, y_n) = (t + y_1, y_2, \dots, y_n), \end{aligned}$$

and this is the desired coordinate description of the flow in the new coordinates.  $\square$

**8.3.16. Higher-order equations.** An ordinary differential equation of order  $k$  (solved with respect to the highest derivative) is an equation



$$(1) \quad y^{(k)} = f(t, y, y', \dots, y^{(k-1)}),$$

where  $f$  is a known function of  $k + 1$  variables,  $t$  is the independent variable, and  $y(t)$  is an unknown function of one variable. This type of equation is always equivalent to a system of  $k$  first-order equations.

Introduce new unknown functions in a variable  $t$  as follows:

$$y_0(t) = y(t), \quad y_1(t) = y'(t), \dots, \quad y_{k-1}(t) = y^{(k-1)}(t).$$

Now, the function  $y(t)$  is a solution of the original equation (1) if and only if it is the first component of the solution of the system of equations

$$\begin{aligned} y_0' &= y_1 \\ y_1' &= y_2 \\ &\vdots \\ y_{k-2}' &= y_{k-1} \\ y_{k-1}' &= f(t, y_0, y_1, \dots, y_{k-1}). \end{aligned}$$

Hence the following direct corollary of the theorems 8.3.9–8.3.12:

SOLUTIONS OF HIGHER-ORDER ODES

**Theorem.** Consider a function  $f(t, y_0, \dots, y_{k-1}) : U \subset \mathbb{R}^{k+1} \rightarrow \mathbb{R}$  with continuous partial derivatives on an open set  $U$ . Then for every point  $(t_0, z_0, \dots, z_{k-1}) \in U$ , there exists a maximal interval  $I_{max} = (x_0 - a, x_0 + b)$ , with positive numbers  $a, b \in \mathbb{R}$ , and a unique function  $y(t) : I_{max} \rightarrow \mathbb{R}$  which is a solution of the  $k$ -th order equation

$$y^{(k)} = f(t, y, y', \dots, y^{(k-1)})$$

with the initial condition

$$y(t_0) = z_0, y'(t_0) = z_1, \dots, y^{(k-1)}(t_0) = z_{k-1}.$$

This solution depends differentiably on the initial conditions and on potential further parameters differentiably entering the function  $f$ . Moreover, the solution is analytic if the latter dependence is analytic.

In particular, the theorem shows that in order to determine unambiguously the solution of an ordinary  $k$ -th order differential equation, the values of the solution and its first  $k - 1$  derivatives must be determined at one point

With a system of  $\ell$  equations of order  $k$ , the same procedure transforms this system to a system of  $k\ell$  first-order equations. Therefore, an analogous statement about existence, uniqueness, continuity, and differentiability is also true.

If the right-hand side  $f$  of the equation is differentiable up to order  $r$  or analytic, including the parameters, than the same property is enjoyed by the solutions as well.

**8.3.17. Linear differential equations.** The operation of differentiation can be viewed as a linear mapping from (sufficiently) smooth functions to functions. Multiplying the derivatives  $(\frac{d}{dt})^j$  of the particular orders  $j$  by fixed functions  $a_j(t)$  and adding these expressions, gives the *linear differential operators*  $y(t) \mapsto D(y)(t)$ :

$$D(y)(t) = a_k(t)y^{(k)}(t) + \dots + a_1(t)y'(t) + a_0(t)y(t).$$

To solve the corresponding *homogeneous linear differential equation* of order  $k$  then means finding a function  $y$  satisfying  $D(y) = 0$ .

The sum of two solutions is again a solution, since for any functions  $y_1$  and  $y_2$ ,

$$D(y_1 + y_2)(t) = D(y_1)(t) + D(y_2)(t).$$

A constant multiple of a solution is again a solution. So the set of all solutions of a  $k$ -th order linear differential equation is a vector space. Apply the previous theorem about existence and uniqueness, to obtain the following:

#### THE SPACE OF SOLUTIONS OF LINEAR EQUATIONS

**Theorem.** *The set of all solutions of a homogeneous linear differential equation of order  $k$  with continuously differentiable coefficients is a vector space of dimension  $k$ . Therefore, the solutions can be described as linear combinations of any set of  $k$  linearly independent solutions. Such solutions are determined uniquely by linearly independent initial conditions on the value of the function  $y(t)$  and its first  $k - 1$  derivatives at a fixed point  $t_0$ .*

**PROOF.** Choose  $k$  linearly independent initial conditions at a fixed point. For each of them, there is a unique solution. A linear combination of these initial condition then leads to the same linear combination of the corresponding solutions. All of the possible initial conditions are exhausted, so the entire space of solutions of the equation is obtained in this way.  $\square$

The same arguments as with the first order linear differential equations in the paragraph 8.3.4 reveal that all solutions of the non-homogeneous  $k$ -th order equation  $D(y) = b(t)$  with a fixed continuous function  $b(t)$  are the sums of one fixed solution  $y(t)$  of this problem and all solutions  $\tilde{y}$  of the corresponding homogeneous equation. Thus the entire space of solutions is an affine  $k$ -dimensional space of functions. The method of variation of constants exploited in 8.3.4 is one of the possible approaches to guess one non-homogeneous solution if we know the complete solution to the homogeneous problem.

We shall illustrate the latter results on the most simple case:

**8.3.18. Linear equations with constant coefficients.** The previous discussion recalls the situation with homogeneous linear difference equations dealt with in paragraph 3.2.1 of the third chapter. The analogy goes further when all of the coefficients  $a_j$  of the differential operator  $D$  are constant. Such first-order equations (1) have solutions as an exponential with an appropriate constant at the argument. Just as in the case of difference equations, it suggests trying whether such a form of the solution  $y(t) = e^{\lambda t}$  with an unknown parameter  $\lambda$  can satisfy an equation of order  $k$ . Substitution yields

$$D(e^{\lambda t}) = (a_k \lambda^k + a_{k-1} \lambda^{k-1} + \cdots + a_1 \lambda + a_0) e^{\lambda t}.$$

The parameter  $\lambda$  leads to a solution of a linear differential equation with constant coefficients if and only if  $\lambda$  is a root of the *characteristic polynomial*  $a_k \lambda^k + \cdots + a_1 \lambda + a_0$ .



If the root  $\lambda = a + ib \in \mathbb{C}$  is not real, then we may consider the complex valued solution  $e^{\lambda t}$ . Since the conjugate root is also involved, we may consider the linear combinations of the solutions  $e^{\lambda t}$  and  $e^{\bar{\lambda}t}$ , providing the real solutions  $e^{at} \sin(bt)$  and  $e^{at} \cos(bt)$ .

If the characteristic polynomial has  $k$  distinct roots, then we have the basis of the whole vector space of solutions. Otherwise, if  $\lambda$  is a multiple root, then direct calculation, making use of the fact that  $\lambda$  is then a root of the derivative of the characteristic polynomial as well, yields that the function  $y(t) = t e^{\lambda t}$  is also a solution. Similarly, for higher multiplicities  $\ell$ , There are  $\ell$  distinct solutions  $e^{\lambda t}, t e^{\lambda t}, \dots, t^{\ell-1} e^{\lambda t}$ .

In the case of a general linear differential equation, a non-zero value of the differential operator  $D$  is wanted. Again, as for systems of linear equations or linear difference equations, the general solution of this type of (non-homogeneous) equations

$$D(y) = b(t),$$

for a fixed function  $b(t)$ , is the sum of an arbitrary solution of this equation and the set of all solutions of the corresponding homogeneous equation  $D(y)(t) = 0$ . The entire space of solutions is a finite-dimensional affine space, hidden in the huge space of functions.

The methods for finding a particular solution are introduced in concrete examples in the other column. In principle, they are based on looking for the solution in a similar form as the right-hand side is, or the method of variation of the constants.

**8.3.19. Matrix systems with constant coefficients.** Before



leaving the area of differential equations, consider a very special case of first-order systems, whose right-hand side is given by multiplication of a matrix  $A \in \text{Mat}_n(\mathbb{R})$  of constant coefficients and an  $n^2$ -dimensional unknown matrix function  $Y(t)$ :

$$(1) \quad Y'(t) = A \cdot Y(t).$$

Clearly this is a strict analogy to the iterative models in chapter 3 and we also met such a system of  $n^2$  equations 8.3.12(2) when preparing the proof of Theorem 8.3.12.

Combine knowledge from linear algebra and univariate function analysis to guess the solution. Define the *exponential of a matrix* by the formula

$$B(t) = e^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k.$$

The right-hand expression can be formally viewed as a matrix whose entries  $b_{ij}$  are infinite series created from the mentioned products. If all entries of  $A$  are estimated by the maximum of their absolute values  $\|A\| = C$ , then the absolute value of the  $k$ -th summand in  $b_{ij}(t)$  is at most  $\frac{|t|^k}{k!} n^k C^{2k}$ . Hence, every series  $b_{ij}(t)$  is necessarily absolutely and uniformly convergent, and it is bound above by the value  $e^{|t|nC^2}$ . Differentiate the terms of the series one by one, to get a uniformly convergent series with limit  $A e^{tA}$ . Therefore, by the

general properties of uniformly convergent series, the derivative

$$\frac{d}{dt}(e^{tA}) = A e^{tA}$$

also equals this expression. The general solution of the system (1) is obtained in the form

$$Y(t) = e^{tA} \cdot Y_0,$$

where  $Y_0 \in \text{Mat}_n(\mathbb{R})$  is the arbitrary initial condition  $Y(0) = Y_0$ . The exponential  $e^{tA}$  is a well defined invertible matrix for all  $t$ . So we have a vector space of the proper dimension, and hence all solutions to the system (1). Notice that in order to get a solution, it is necessary to multiply by  $Y_0$  from the right.

It is remarkable that dealing with a vector equation with a constant matrix  $A \in \text{Mat}_n(\mathbb{R})$ ,

$$(2) \quad y' = A \cdot y,$$

for an unknown function  $y : \mathbb{R} \rightarrow \mathbb{R}^n$ , then the columns of the matrix exponential  $e^{tA}$  provide  $n$  linearly independent solutions. The general solution is then given by linear combinations of them.

The general solutions of the system (2) may be understood much better by invoking some linear algebra – the Jordan canonical form of linear mappings, see e.g. 3.4.10. In terms of vector fields  $X$ , the system has the linear expression  $X(y) = \Phi(y)$  where  $\Phi$  is the linear mapping with the matrix  $A$  in coordinates. Clearly linear transformations of the system lead to another vector field with such linear description, since the differential of a linear mapping is the mapping itself.

Any linear transformation of coordinates with the (constant) matrix  $T$  transforms the system into

$$\tilde{y}' = (Ty)' = (TAT^{-1}) \cdot (Ty) = \tilde{A} \cdot \tilde{y}.$$

In particular, a suitable change of coordinates  $T$  provides the matrix  $\tilde{A}$  in the Jordan canonical form expressing  $\Phi$  as a sum of two commuting linear mappings  $\Phi = \Phi_d + \Phi_n$  with  $\Phi_d$  diagonalisable and  $\Phi_n$  nilpotent. Moreover, the decomposition of the nilpotent part into the sum of cyclic nilpotent mappings provides the Jordan blocks

$$\begin{aligned} J_\lambda &= \begin{pmatrix} \lambda & 1 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix} = (J_\lambda)_d + (J_\lambda)_n \\ &= \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{pmatrix} + \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \end{aligned}$$

Splitting the system (2) into block-wise diagonal form splits also the space of the solutions generated by the exponential  $e^{tA}$  into the corresponding blocks (all the powers  $A^k$  enjoy the same block structure). So we can work with the matrix  $A$  already in the form on one such block  $J_\lambda = (J_\lambda)_d + (J_\lambda)_n$ . But for any two commuting matrices  $C$  and  $D$ , the exponentials  $e^{tC}$  and  $e^{tD}$  commute. So the exponential  $e^{tD}$  of the

nilpotent  $D = (J_\lambda)_n$  can be computed as the finite sum

$$E + t \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} + \cdots + \frac{t^{k-1}}{(k-1)!} \begin{pmatrix} 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

where  $k$  is the size of the block and  $E$  is the identity matrix. The solution of the corresponding matrix system is

$$Y(t) = e^{\lambda t E} \cdot e^{Dt} = e^{\lambda t} e^{Dt} = \begin{pmatrix} e^{\lambda t} & t e^{\lambda t} & \cdots & \frac{t^{k-1}}{(k-1)!} e^{\lambda t} \\ 0 & e^{\lambda t} & \cdots & \frac{t^{k-2}}{(k-2)!} e^{\lambda t} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda t} \end{pmatrix}$$

Finally,  $k$  independent solutions can be written down by inspecting the individual columns in  $Y(t)$ . Notice that the canonical basis  $(e_1, \dots, e_k)$  provides just the chain of vectors with  $D(e_k) = e_{k-1}$ ,  $k = 2, \dots, k$ , while  $D(e_1) = 0$ . Now the  $k$  independent solutions are:

$$\begin{aligned} y_1(t) &= e^{\lambda t} e_1 \\ y_2(t) &= e^{\lambda t} (t e_1 + e_2) \\ &\vdots \\ y_k(t) &= e^{\lambda t} \left( \frac{t^{k-1}}{(k-1)!} e_1 + \frac{t^{k-2}}{(k-2)!} e_2 + \cdots + t e_{k-1} + e_k \right). \end{aligned}$$

The result can be easily transferred back to the original coordinates in which the system (2) was given. Finding the decomposition of the space into Jordan blocks and finding the chains of basis vectors  $v_i$  realizing the cyclic nilpotent components, we arrive at the independent solutions by replacing the  $e_i$  by  $v_i$ .

The findings are summarised in the following theorem, one of many attributed to Euler.

**Theorem.** *All solutions of the system (2) are linear combinations of those in the form combining exponential and polynomial expressions*

$$y(t) = e^{\lambda t} \sum_{j=0}^k p_j t^j$$

where  $k$  is the order of nilpotency of the Jordan block corresponding to the eigenvalue  $\lambda$  of the matrix  $A$ ,  $p_j$  are suitable constant vectors. In particular, if the nilpotent part of  $A$  is trivial, then  $k = 0$ .

This important result allows many generalizations. For example, the Floquet-Lyapunov theory generalizes this behaviour of solutions to systems with periodically time-dependent matrices  $A(t)$ .

**8.3.20. Return to singular points.** Finally, recall the first-order matrix system in paragraph 8.3.12 when the derivative of the solutions of vector equations with respect to the initial conditions was discussed. Consider a differentiable vector field  $X(x)$  defined on a neighbourhood of its singular point  $x_0 \in \mathbb{R}^n$ , i.e.  $X(x_0) = 0$ . Then, the point  $x_0$  is a fixed point of its flow  $\text{Fl}_t^X(x)$ .

The differential  $\Phi(t) = D_x \text{Fl}_t^X(x_0)$  satisfies the matrix system with initial condition (see (2) on page 778)

$$\Phi'(t) = D^1 X(x_0) \cdot \Phi(t), \quad \Phi(0) = E.$$

The important point is that the differential  $D^1 X$  is evaluated along the constant flow line  $x_0$ , since this is a singular point of the system.

The solution is known explicitly, and this describes the evolution of the differential  $\Phi(t)$  of the vector field's flow at the singular point  $x_0$ ,

$$\Phi(t) = e^{tA}, \quad A = D^1 X(x_0).$$

This is a useful step in analysing the qualitative behaviour in a neighbourhood of the stationary point  $x_0$ .

Consider the Lotka-Volterra system from this point of view. Use the coordinates  $(x, y)$  and parameters  $\alpha, \beta, \gamma, \delta$  exactly as in 8.3.10. In particular all these quantities are assumed to be positive.

The vector field in question is

$$X(x, y) = (x(\alpha - \beta y), y(-\gamma + \beta \delta x)).$$

So there is a single singular point (except  $(0, 0)$ )

$$(x_0, y_0) = \left( \frac{\gamma}{\delta \beta}, \frac{\alpha}{\beta} \right)$$

and the differential of  $X$  at this point is

$$D^1 X(x_0, y_0) = \begin{pmatrix} \alpha - \beta y_0 & -\beta x_0 \\ \delta \beta y_0 & \delta \beta x_0 - \gamma \end{pmatrix} = \begin{pmatrix} 0 & -\frac{\gamma}{\delta} \\ \alpha \delta & 0 \end{pmatrix}$$

The characteristic polynomial of the latter matrix is  $\lambda^2 + \gamma\alpha$  and so there are two complex conjugated roots  $\lambda = \pm i\sqrt{\alpha\gamma}$ . As is known from linear algebra, such a matrix describes a rotation in suitable coordinates. Compute the real and imaginary components of the eigenvectors corresponding to  $\lambda$  (as developed in linear algebra), to obtain the matrix solution in the form

$$Y(t) = \begin{pmatrix} \cos \sqrt{\alpha\gamma} t & -\sin \sqrt{\alpha\gamma} t \\ \delta \sqrt{\frac{\alpha}{\gamma}} \sin \sqrt{\alpha\gamma} t & \delta \sqrt{\frac{\alpha}{\gamma}} \cos \sqrt{\alpha\gamma} t \end{pmatrix}.$$

The columns are the two independent solutions  $y(t)$  (they differ just by the phase of the linearly distorted rotation).

This might be useful information for further analysis of the model around its singular point. For example, the parameter  $\beta$  does not appear explicitly here, while  $\delta$  influences the distortion of the flow lines from being circles. Compare this with the illustrations on page 774. For a first approximation, guess that the sizes of the populations of both the prey and the predator oscillate regularly around the values of the singular point if the initial conditions are near this point.

**8.3.21. A note about Markov chains.** In the third chapter, we dealt with iterative processes, where the stochastic matrices and Markov processes determined by them played an interesting role. Recall that a matrix  $A$  is stochastic if the sum of each of its columns is one. In other words,



$$(1 \dots 1) \cdot A = (1 \dots 1).$$

Take the exponential  $e^{tA}$  to obtain

$$(1 \dots 1) \cdot e^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!} (1 \dots 1) \cdot A^k = e^t (1 \dots 1).$$

Therefore, for every  $t$ , the invertible matrix

$$B(t) = e^{-t} e^{tA}$$

is stochastic, if  $A$  is stochastic. Add stochastic initial conditions  $B_0$ , to get the flow  $B(t) = e^{-t} e^{tA} \cdot B_0$ , which is a continuous version of the Markov process (infinitesimally) generated by the stochastic matrix  $A$ .

Differentiating with respect to  $t$ , yields

$$\frac{d}{dt} B(t) = -e^{-t} e^{tA} \cdot B_0 + e^{-t} A e^{tA} \cdot B_0 = (-E + A)B(t),$$

so the matrix  $B(t)$  is the solution of the matrix system of equations with constant coefficients

$$Y'(t) = (A - E) \cdot Y(t)$$

with the stochastic matrix  $A$ .

This can be explained intuitively. If the matrix  $A$  is stochastic, then the instantaneous increment of the stochastic vector  $y(t)$  in the vector system with the matrix  $A$ ,  $y'(t) = A \cdot y(t)$ , is again a stochastic vector. However, it is desired that the Markov process keeps the vector  $y(t)$  stochastic for all  $t$ . Hence, the sum of increments of the particular components of the vector  $y(t)$  must be zero, which is guaranteed by subtracting the identity matrix.

As seen above, the columns of the matrix solution  $Y'(t)$  create a basis of all solutions  $y'(t)$  of the vector system.

Much information can be obtained about the solutions by using some linear algebra. For example, suppose that the matrix  $A$  is primitive, that is, suppose one of its powers has only positive entries, see 3.3.3 on page 214. Then its powers converge to a matrix  $A_\infty$ , all of whose columns are eigenvectors corresponding to the eigenvalue 1.

Next, estimate the difference between the solution  $Y'(t)$  for large  $t$  and the constant matrix  $A_\infty$ . There are two consequences from the latter convergence. First, there exists a universal constant bound for all powers  $\|A^k - A_\infty\| \leq C$ . Second, for every small positive  $\varepsilon$ , there is an  $N \in \mathbb{N}$  such that for all  $k \geq N$ ,  $\|A^k - A_\infty\| \leq \varepsilon$ . Hence,

$$\begin{aligned} \left\| e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k - e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} A_\infty \right\| \\ \leq e^{-t} \sum_{k < N} \frac{t^k}{k!} C \|A_\infty\| + \varepsilon \|A_\infty\|. \end{aligned}$$

Let  $t \rightarrow \infty$ . The limit of the expression  $f(t) = e^{-t} \sum_{k < N} \frac{t^k}{k!}$  can easily be computed by iterative application of l'Hospital's rule. Differentiation of the sum yields the same, only for  $N$  smaller by one, and the derivative in the denominator is not changed, so the limit is zero. Therefore, for fixed  $\varepsilon$ , there is a time  $T$  such that  $f(t)$  would be less than  $\varepsilon$  for  $t \geq T$ . The whole expression has

been estimated (for  $n \geq N$  and  $t \geq T > 0$ ) by the value  $\varepsilon(C + 1)\|A_\infty\|$ .

Summarizing, a very interesting statement is proved, which resembles the discrete version of Markov processes:

#### CONTINUOUS PROCESSES WITH A STOCHASTIC MATRIX

**Theorem.** Every primitive stochastic matrix  $A$  determines a vector system of equations

$$y' = (A - E) \cdot y$$

with the following properties:

- the basis of the vector space of all solutions is given by the columns of the stochastic matrix

$$Y(t) = e^{-t} e^{tA},$$

- if the initial condition  $y_0 = y(t_0)$  is a stochastic vector, then the solution  $y(t)$  is also a stochastic vector for all values of  $t$ ,
- every stochastic solution converges for  $t \rightarrow \infty$  to the stochastic eigenvector  $y_\infty$  of the matrix  $A$  corresponding to the eigenvalue 1.

**8.3.22. Remarks on numerical methods.** Except for the exceptionally simple equations, for example, linear equations with constant coefficients, analytically solvable equations are seldom encountered in practice. Therefore, some techniques are required to approximate the solutions of the equations.

Approximations have already been considered in many other situations. (Recall the interpolation polynomials and splines, exploitation of Taylor polynomials in methods for numerical differentiation and integration, Fourier series etc.). With a little courage, consider difference and differential equations to be mutual approximations. In one direction, replace differences with differentials (for example, in economical or population models). For other situations the differences may imitate well continuous changes in models.

Use the terminology for asymptotic estimates, as introduced in 6.1.12. In particular, an expression  $G(h)$  is asymptotically equal to  $F(h)$  for  $h$  approaching zero or infinity, and write  $G(h) = O(F(h))$ , if the finite limit of  $G(h)/F(h)$  exists.

A good example is the approximation of a multivariate function  $f(x)$  by its Taylor polynomial of order  $k$  at a point  $x_0$ . Taylor's theorem says that the error of this approximation is  $O(\|h\|^{k+1})$ , where  $h$  is the increment of the argument  $h = x - x_0$ .

In the case of ordinary differential equations, the simplest scheme is approximation using *Euler polygons*. Present this method for a single ordinary equation with two quantities: one independent and one dependent. It works analogously for systems of equations where scalar quantities and their derivatives in time  $t$  are replaced with vectors dependent on time and their derivatives. This procedure was used before in the proof of the Peano's existence theorem, see 8.3.8.

Consider an equation

$$y' = f(t, y)$$

with continuous right-hand  $f$ . Denote the discrete increment of time by  $h$ , i.e. set  $t_n = t_0 + nh$ . It is desired to approximate  $y(t)$ . It follows from Taylor's theorem (with remainder of order two) and the equation that

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + y'(t_n)h + O(h^2) \\ &= y(t_n) + f(t_n, y(t_n))h + O(h^2). \end{aligned}$$

Define recurrently the values  $y_j$  by the first order formula

$$y_{j+1} = y_j + f(t_j, y_j)h.$$

This leads to the local approximation error  $O(h^2)$ , occurring in one step of the recurrence.

If  $n$  such steps are needed with increment  $h$  from  $t_0$  to  $t = t_n$ , the error could be up to  $nO(h^2) = \frac{1}{h}(t-t_0)O(h^2) = O(h)$ . More care is needed, since the function  $f(t, y)$  is evaluated in the approximate points  $(t_i, y_i)$  and the already approximate previous values  $y_j$ . In order to keep control,  $f(t, y)$  must be Lipschitz in  $y$ . Assuming inductively that the estimate is true for all  $i < j$ ,

$$|f(t_j, y(t_j)) - f(t_j, y_j)| \leq C|y(t_j) - y_j| \leq C|t - t_0|O(h)$$

where  $C$  is the Lipschitz constant, assuming that the error does not exceed  $O(h)$  with globally valid constant for  $y_j$ . Inductively, the expected bound  $O(h)$  for the global error estimate is obtained. Think about the details.

The Euler procedure is the simplest method within the class of the *Runge-Kutta* methods.

Dealing with higher order equations, we may either view them as vector valued first order systems (as in the theoretical column) and then even Euler method provides results for the initial condition on the necessary number of derivatives in one point. But in practical problems, it is often needed to find solutions passing through more than one prescribed point. For example, with second order equations, prescribe two values  $y(t_1)$  and  $y(t_2)$  of the solution. This would need completely different methods.

**O. Additional exercises to the whole chapter**

**8.O.1.** A basin with volume 300 hl contains 100 hl of water in which 50 kg of salt is dissolved. Water with 2 kg of salt per 1 hl starts flowing into the basin at 6 hl/min. The mixture, being kept homogeneous by permanent stirring, leaves the basin at 4 hl/min. Express the amount of salt (in kg) in the basin after  $t$  minutes have expired as a function of the variable  $t \in [0, 100]$ .

**8.O.2.** During a controlled experiment, a small smelting furnace is slowly cooling down while the outer temperature keeps at 300 K. The experiment began at noon. At 1 pm, the temperature in the furnace was estimated at 1300 K. At 3 pm, it was only 550 K. Supposing the measurements were accurate, compute what the temperature in the furnace was at 2 pm.

**8.O.3.** The half-life of the radioactive sulfur isotope  $^{35}\text{S}$  is 87.5 days. After what period are there only 900 grams left of the original amount of 1 kilogram of this isotope? (You may express the result in terms of the natural logarithm.)

**8.O.4.** The half-time of a radioactive element  $A$  is 5 years; for an element  $B$ , it is 1 year. If we have 5 kg of element  $B$  and 1 kg of element  $A$ , after what period will we have the same amount of both? (You may express the result in terms of the natural logarithm.)

**8.O.5.** The half-time of a radioactive element  $A$  is 8 years; for an element  $B$ , it is 2 years. If we have 3 kg of element  $B$  and 1 kg of element  $A$ , after what period will we have the same amount of both? (You may express the result in terms of the natural logarithm.)

**8.O.6.** The half-life of the radioactive cobalt isotope  $^{60}\text{Co}$  is 5.27 years. Having 4 kg of this isotope, after what period does 1 kg of it decay? (You may express the result in terms of the natural logarithm.)

**8.O.7.** Solve the following differential equation for the function  $y = y(x)$ :

$$y' = \frac{1 + y^2}{1 + x^2}.$$

**8.O.8.** Determine all solutions of the following equation with separated variables:

$$y - y^2 + xy' = 0.$$

**8.O.9.** Solve the equation

$$1 + \frac{dy}{dx} = e^y.$$

**8.O.10.** Solve the equation  $2y = x^3y'$ .

**8.O.11.** Determine all solutions of the equation

$$\sqrt{4 - y^2} dx + y dy = 0.$$

**8.O.12.** Solve

$$y' \tan x = y^2 + 1 - 2y.$$

**8.O.13.** Determine the general solution of the differential equation

$$\frac{x^2+1}{x} = \frac{y}{1-y^2} y'.$$

**8.O.14.** Find the general solution of the differential equation

$$(x + 1) dy + xy dx = 0.$$



**8.O.15.** Find the solution of the differential equation

$$\sin y \cos x \, dy = \cos y \sin x \, dx$$

which satisfies  $4y(0) = \pi$ .

**8.O.16.** Solve the initial problem

$$(x^2 + 1)(y^2 - 1) + xy y' = 0, \quad y(1) = \sqrt{2}.$$

**8.O.17.** Determine the particular solution of the equation

$$y' \sin x = y \ln y$$

which goes through the point  $[\pi/2, e]$ .

**8.O.18.** Find all solutions of the differential equation

$$2(1 + e^x)yy' = e^x,$$

which satisfy  $y(0) = 0$ .

**8.O.19.** Solve the homogeneous equation

$$(xy' - y) \cos \frac{y}{x} = x.$$

**8.O.20.** Determine the general solution of the homogeneous differential equation  $y^3 = x^3 y'$ .

**8.O.21.** Find all solutions of the equation

$$xy' = \sqrt{x^2 - y^2} + y.$$

**8.O.22.** Determine the general solution if we are given

$$xy' = y \cos \left( \ln \frac{y}{x} \right).$$

**8.O.23.** Solve the equation  $(x + y) \, dx - (x - y) \, dy = 0$  as homogeneous.

**8.O.24.** Calculate  $y' = (x + y)^2$ .

**8.O.25.** Find the general solution for

$$y' = \frac{x-y+3}{x+y-5}.$$

**8.O.26.** Calculate

$$y' = \frac{x-y+1}{x-y}.$$

**8.O.27.** Determine all solutions of the differential equation

$$y' = \frac{5y-5x-1}{2y-2x-1}.$$

**8.O.28.** Find the general solution of the equation

$$y' = \frac{x-y-1}{x+y+3}.$$

8.0.29. Determine the general solution for the equation

$$y' = \frac{2x-y-5}{x-3y-5}.$$

8.0.30. Express the solutions of the equation

$$y' = \frac{x+2y-7}{x-3}$$

as explicitly given functions.

8.0.31. Using the method of constant variation, calculate  $y' + 2y = x$ .

8.0.32. Determine the general solution of the equation  $y' = 6x + 2y + 3$ .

8.0.33. Solve the linear equation

$$y' = 4xy + (2x + 1)e^{2x^2}.$$

8.0.34. Solve the equation  $y'x + y = x \ln x$ .

8.0.35. Calculate the linear differential equation

$$y'x = y + x^2 \ln x.$$

8.0.36. Find all solutions of the equation

$$y' \cos x = (y + 2 \cos x) \sin x.$$

8.0.37. Find the solution of the equation  $y' = 6x - 2y$  which satisfies the initial condition  $y(0) = 0$ .

8.0.38. Solve the initial problem

$$y' + y \sin x = \sin x, \quad y\left(\frac{\pi}{2}\right) = 2.$$

8.0.39. Find the solution of the equation  $y' = 4y + \cos x$  which goes through the point  $[0, 1]$ .

8.0.40. Solve the following equation for any  $a, b \in \mathbb{R}$ :

$$xy' + y = e^x, \quad y(a) = b.$$

8.0.41. Determine the general solution of the equation

$$3x^2y' + xy = \frac{1}{y^2}.$$

8.0.42. Solve the Bernoulli equation

$$y' = xy - y^3 e^{-x^2}.$$

8.0.43. Calculate the Bernoulli equation

$$y' - \frac{y}{x} = y^2 \sin x.$$

8.0.44. Find all solutions of the equation

$$y' = \frac{4y}{x} + x\sqrt{y}.$$

8.0.45. Solve the equation

$$xy' + 2y + x^5y^3e^x = 0.$$

8.0.46. Find the general solution of the following equation provided  $a, b > 0$ :

$$y dy = \left( a \frac{y^2}{x^2} + b \frac{1}{x^2} \right) dx.$$

8.0.47. Interchanging the variables, solve

$$2y + (y^2 - 6x)y' = 0.$$

8.0.48. Solve the equation

$$y' = \frac{y}{2y \ln y + y - x}.$$

8.0.49. Calculate the general solution of the following equation:

$$x dx = \left( \frac{x^2}{y} - y^3 \right) dy.$$

8.0.50. Interchanging the variables, calculate

$$(x + y) dy = y dx + y \ln y dy.$$

8.0.51. Solve

$$y' (e^{-y} - x) = 1.$$

8.0.52. Calculate

$$y' = \frac{1}{2x - y^2}.$$

8.0.53. Solve the equation

$$2y dx + x dy = 2y^3 dy.$$

8.0.54. Calculate

$$y'' + 3y' + 2y = (20x + 29)e^{3x}.$$

8.0.55. Find any solution of the non-homogeneous linear equation

$$y'' + y' + \frac{5}{2}y = 25 \cos(2x).$$

**8.0.56.** Determine the solution of the equation

$$y'' + 2y' + 2y = 3e^{-x} \cos x.$$

**8.0.57.** Find the solution of the equation

$$y'' = 2y' + y + 1,$$

which satisfies  $y(0) = 0$  and  $y'(0) = 1$ .

**8.0.58.** Find the solution of the equation

$$y'' = 4y - 3y' + 1$$

which satisfies  $y(0) = 0$  and  $y'(0) = 2$ .

**8.0.59.** Determine the general solution of the linear equation

$$y'' - 2y' + 5y = 5e^{2x} \sin x.$$

**8.0.60.** Taking advantage of the special form of the right-hand side, find all solutions of the equation

$$y'' + y' = x^2 - x + 6e^{2x}.$$

**8.0.61.** Solve

$$y^{(4)} - 2y'' + y = 8(e^x + e^{-x}) + 4(\sin x + \cos x).$$

**8.0.62.** Using the method of constant variation, calculate

$$y'' - 2y' + y = \frac{e^x}{x}.$$

**8.0.63.** Solve

$$y'' + 4y' + 4y = e^{-2x} \ln x.$$

**8.0.64.** Using the method of constant variation, find the general solution of the equation

$$y'' + 4y = \frac{1}{\sin(2x)}.$$

**8.0.65.** Solve the equation  $y'' + y = \tan^2 x$ .

**8.0.66.** Find the solution of the differential equation

$$y^{(3)} = -2y'' - 2y' - y + \sin(x)$$

which satisfies  $y(0) = -\frac{1}{2}$ ,  $y'(0) = \frac{\sqrt{3}}{2}$ , and  $y''(0) = -1 - \frac{\sqrt{3}}{2}$ .

**8.0.67.** Calculate the equation  $y''' - 2y'' - y' + 2y = 0$ .

**8.0.68.** Find the general solution of the equation

$$y^{(4)} + 2y'' + y = 0.$$

**8.O.69.** Solve

$$y^{(6)} + 2y^{(5)} + 4y^{(4)} + 4y''' + 5y'' + 2y' + 2y = 0.$$

**8.O.70.** Find the general solution of the linear equation

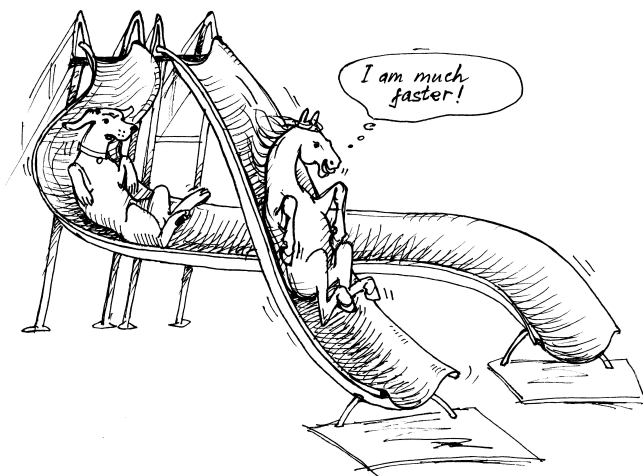
$$y^{(5)} - 3y^{(4)} + 2y''' = 8x - 12.$$



**Key to the exercises**

## Continuous models – further selected topics

*The World might be discrete in reality.  
But the continuous models are useful anyhow ...*



### A. Exterior differential calculus

### B. Applications of Stoke's theorem

9.B.1. Compute

$$\int_c (x - y)dx + x dy,$$

where  $c$  is the positively oriented curve represented by the perimeter of the square  $ABCD$  with vertices  $A = [2, 2]$ ;  $B = [-2, 2]$ ;  $C = [-2, -2]$ ;  $D = [2, -2]$ .

**Solution.** Using Green's theorem (see ??), we reduce the given curve integral to an area (multiple) integral. The integral is of the form  $\int_c f(x, y) dx + g(x, y) dy$ , where  $f(x, y) = x - y$  and  $g(x, y) = x$ . The needed partial derivatives of the functions  $f(x, y)$  and  $g(x, y)$  are thus  $f_y(x, y) = -1$  and  $g_x(x, y) = 1$ . All of the functions  $f(x, y)$ ,  $g(x, y)$ ,  $f_y(x, y)$ , and  $g_x(x, y)$  are continuous on  $\mathbb{R}^2$ , so we can use Green's theorem:

$$\int_c (x - y)dx + xdy = \iint_D (1 + 1)dxdy = 2 \iint_D dxdy$$

This chapter presents several glimpses towards more serious applications of the differential and integral calculus. We cannot be ambitious in covering the displayed topics extensively. Thus, after reasonably detailed introduction to more geometric approach to the differential and integral calculus, we present rather quick surveys and comments on partial differential equations, variational calculus, and complex analysis. We hope the readers will get excited at least by some of them and look for further resources themselves.

many pictures missing!!!

### 1. Exterior differential calculus and integration

We have already seen how to optimize functions on subsets in  $\mathbb{R}^n$ , but how to integrate quantities over such domains? For example, if we have 2-dimensional membrane in  $\mathbb{R}^3$  and we know the infinitesimal flow of some liquid through it, how to compute how much went through within a time interval?

In order to understand such questions properly, we formalize the concept of the level sets  $M_b$  from the paragraph 8.1.24 devoted to the implicit functions and provide a geometric explanation of the integration process. Then we quite easily arrive at several powerful tools, including the Stokes theorem, which is a higher dimensional extension to the fundamental theorem of (univariate) calculus, and the Frobenius theorem which generalizes the integration of prescribed line elements into a solution of an ODE to higher dimensions.

#### 9.1.1. Vector fields and differential forms.



Let us come back to the concept of tangent vectors and vector fields, cf. 8.3.14 where we introduced the *tangent space*  $TU = U \times \mathbb{R}^n$  as the set of all possible tangent vectors at the points of an open subset  $U \subset \mathbb{R}^n$ . There is the projection  $p : TU \rightarrow U$  assigning the foot points to the tangent vectors, we write  $T_x U$  for the vector space of all vectors  $X$  with  $p(X) = x$  at a point  $x \in U$ , and we use the notation  $\mathcal{X}(U)$  for the set of all smooth vector fields on the open subset  $U$ .

The linear combinations of the special vector fields  $\frac{\partial}{\partial x_i}$  admitting smooth functions as the coefficients generate the entire  $\mathcal{X}(U)$ . Thus we write general vector fields as

$$X(x) = X_1(x) \frac{\partial}{\partial x_1} + \dots + X_n(x) \frac{\partial}{\partial x_n}.$$

$$= 2 \int_{-2}^2 \int_{-2}^2 dx dy = 2[x]_{-2}^2 \cdot [y]_{-2}^2 = 32.$$

**9.B.2.** Compute

$$\int_c x^4 dx + xy dy,$$

where  $c$  is the positively oriented curve going through the vertices  $A = [0, 0]$ ;  $B = [1, 0]$ ;  $C = [0, 1]$ .

**Solution.** The curve  $c$  is the boundary of the triangle  $ABC$ . The integrated functions are continuously differentiable on the whole  $\mathbb{R}^2$ , so we can use Green's theorem:

$$\begin{aligned} \int_c x^4 dx + xy dy &= \iint_D y dx dy = \int_0^1 \int_0^{1-x} y dx dy \\ &= \int_0^1 \left[ \frac{y^2}{2} \right]_0^{1-x} dx = \int_0^1 \left[ \frac{x^2 - 2x + 1}{2} \right] dx \\ &= \frac{1}{2} \left[ \frac{x^3}{3} - \frac{2x^2}{2} + x \right]_0^1 = \frac{1}{6}. \end{aligned}$$

**9.B.3.** Calculate

$$\int_c (xy + x + y) dx + (xy + x - y) dy,$$

where  $c$  is the circle with radius 1 centered at the origin.

**Solution.** Again, the prerequisites of Green's theorem are satisfied, so we can use Green's theorem, which now gives

$$\begin{aligned} &\int_c (xy + x + y) dx + (xy + x - y) dy \\ &= \iint_D y + 1 - x - 1 dx dy \\ &= \int_0^1 \int_0^{2\pi} r^2 (\sin \varphi - \cos \varphi) dr d\varphi \\ &= \int_0^1 r^2 dr \int_0^{2\pi} \sin \varphi - \cos \varphi d\varphi \\ &= \frac{1}{3} [-\cos \varphi - \sin \varphi]_0^{2\pi} = 0. \end{aligned}$$

**9.B.4.** Compute  $\int_c (2e^{2x} \sin y - 3y^3) dx + (e^{2x} \cos y + \frac{4}{3}x^3) dy$ , where  $c$  is the positively oriented ellipse  $4x^2 + 9y^2 = 36$ .

**Solution.:** We will use Green's theorem, choosing the linear deformation of polar coordinates  $x = 3r \cos \varphi$ ,  $\varphi \in [0, 2\pi]$ ,

As we know, every differentiable mapping  $F : U \rightarrow V$  between two open sets  $U \subset \mathbb{R}^n$ ,  $V \subset \mathbb{R}^m$  defines the mapping  $F_* : TU \rightarrow TV$  by applying the differential  $D^1F$  to the individual tangent vectors. Thus if  $y = F(x) = (f_1(x), \dots, f_m(x))$  then  $F_* : T_x U \rightarrow T_{F(x)} V$ ,

$$F_* \left( \sum_{i=1}^n X_i(x) \frac{\partial}{\partial x_i} \right) (y) = \sum_{j=1}^m \left( \sum_{i=1}^n \frac{\partial f_j(x)}{\partial x_i} X_i(x) \right) \frac{\partial}{\partial y_j} (y).$$

When we studied the vector spaces in chapter two, we came across the useful concept of linear forms. They were defined in paragraph 2.3.17 on page 122. This idea extends naturally now. A scalar valued linear mapping defined on the tangent space  $T_x U$  is such a linear form at the foot point  $x$ . The vector space of all such forms  $T_x^* U = (T_x U)^*$  is thus naturally isomorphic to  $\mathbb{R}^{n*}$  and the collection  $T^* U$  of these spaces comes equipped by the projection to the foot points, let us denote it again by  $p$ . Having a mapping  $\eta : U \subset \mathbb{R}^n \rightarrow T^* U$  with values  $\eta(x) \in T_x^* U$  on an open subset  $U$ , i.e.,  $p \circ \eta = \text{id}_U$ , we talk about a *differential form*  $\eta$  on  $U$ , or a *linear form*.

Every differentiable function  $f$  on an open subset  $U \subset \mathbb{R}^n$  defines the differential form  $df$  on  $U$  (cf. 8.1.7). We use the notation  $\Omega^1(U)$  for the set of all smooth linear differential forms on the open set  $U$ .

In the chosen coordinates  $(x_1, \dots, x_n)$  we can use the differentials of the particular coordinate functions to express every linear form  $\eta$  as

$$\eta(x) = \eta_1(x) dx_1 + \dots + \eta_n(x) dx_n,$$

where  $\eta_i(x)$  are uniquely determined functions. Such a form  $\eta$  evaluates on a vector field  $X(x) = X_1(x) \frac{\partial}{\partial x_1} + \dots + X_n(x) \frac{\partial}{\partial x_n}$  as

$$\eta(X)(x) = \eta(x)(X(x)) = \eta_1(x) X_1(x) + \dots + \eta_n(x) X_n(x).$$

If the form  $\eta$  is the differential of a function  $f$ , we get just back the expression

$$X(f)(x) = df(X(x)) = \frac{\partial f}{\partial x_1} X_1(x) + \dots + \frac{\partial f}{\partial x_n} X_n(x)$$

for the derivative of  $f$  in the direction of the vector field  $X$ .

**9.1.2. Exterior differential forms.** As we discussed already in chapters 1 and 4, the volume of  $k$ -dimensional parallelepipeds  $S$ , as a quantity depending of the  $k$  vectors spanning  $S$ , is an antisymmetric  $k$ -linear form on the vectors, see 2.3.23 on page 127. Remember also the computation of the volume of parallelepipeds in terms of determinants in 4.1.19 on page 334.

Thus, if we want to talk about the (linearized) volume on  $k$ -dimensional objects, we need a concept which will be linear in  $k$  distinct tangent vector arguments and will assign a scalar quantity to them. Moreover, we will require that interchanging any pair of arguments swaps the sign, in accordance with the orientations.



$y = 2r \sin \varphi$   $r \in [0, 1]$ ,  
 leading to (the Jacobian of the transformation is  $6r$ ):

$$\begin{aligned} & \int_c (2e^{2x} \sin y - 3y^3) dx + (e^{2x} \cos y + \frac{4}{3}x^3) dy = \\ & = \iint_D 2e^{2x} \cos y + 4x^2 - (2e^{2x} \cos y - 9y^2) dx dy = \\ & = \int_0^1 \int_0^{2\pi} 6r [4(3r \cos \varphi)^2 + 9(2r \sin \varphi)^2] = \\ & = 216 \int_0^1 r^3 dr \int_0^{2\pi} d\varphi = 216 \cdot [\frac{r^4}{4}]_0^1 \cdot 2\pi = 108\pi. \end{aligned}$$

**9.B.5.** Compute

$$\int_c (e^x \ln y - y^2 x) dx + \left( \frac{e^x}{y} - \frac{1}{2}x^2 y \right) dy,$$

where  $c$  is the positively oriented circle  $(x-2)^2 + (y-2)^2 = 1$ .

**Solution.**

$$\begin{aligned} & \int_c (e^x \ln y - y^2 x) dx + \left( \frac{e^x}{y} - \frac{1}{2}x^2 y \right) dy = \\ & = \iint_D \frac{e^x}{y} - xy - \frac{e^x}{y} + 2xy dx dy = \\ & = \int_0^1 \int_0^{2\pi} r(r \cos \varphi + 2) \cdot (r \sin \varphi + 2) dr d\varphi = \\ & = \int_0^1 \int_0^{2\pi} r^3 \sin \varphi \cos \varphi + 2r^2(\sin \varphi + \cos \varphi) + 4r dr d\varphi = \\ & = \frac{1}{4} \int_0^{2\pi} \sin \varphi \cos \varphi d\varphi + \frac{2}{3} \int_0^{2\pi} \sin \varphi + \cos \varphi d\varphi + 4\pi = \\ & = \frac{1}{3} \left[ \frac{\sin^2 \varphi}{2} \right]_0^{2\pi} + [-\cos \varphi + \sin \varphi]_0^{2\pi} + 4\pi = 4\pi. \end{aligned}$$

**9.B.6.** Calculate the integral

$$\int_c (e^x \sin y - xy^2) dx + \left( e^x \cos y - \frac{1}{2}x^2 y \right) dy,$$

where  $c$  is the positively oriented circle  $x^2 + y^2 + 4x + 4y + 7 = 0$ .

**9.B.7.** Compute

$$\int_c (3y - e^{\sin x}) dx + (7x + \sqrt{y^4 + 1}) dy,$$

**Definition.** The vector space of all  $k$ -linear antisymmetric forms on a tangent space  $T_x U$ ,  $U \subset \mathbb{R}^n$ , will be denoted by  $\Lambda^k(T_x U)^*$ . We talk about *exterior  $k$ -forms* at the point  $x \in U$ .

The assignment of a  $k$ -form  $\eta(x) \in \Lambda^k T_x^* U$  to every point  $x \in U$  from an open subset in  $\mathbb{R}^n$  defines an *exterior differential  $k$ -form* on  $U$ . The set of smooth exterior  $k$ -forms on  $U$  is denoted  $\Omega^k(U)$ .

Next, let us consider a smooth mapping  $G : V \rightarrow U$  between two open sets  $V \subset \mathbb{R}^m$  and  $U \subset \mathbb{R}^n$ , an exterior  $k$ -form  $\eta(G(x)) \in \Lambda^k(T_{G(x)}^* U)$ , and choose arbitrarily  $k$  vectors  $X_1(x), \dots, X_k(x)$  in the tangent space  $T_x V$ . Just like in the case of linear forms, we can evaluate the form  $\eta$  at the images of the vectors  $X_i$  using the mapping  $y = G(x) = (g_1(x), \dots, g_n(x))$ . This operation is called the *pullback of the form  $\eta$*  by  $G$ .

$$\begin{aligned} G^*(\eta(G(x)))(X_1(x), \dots, X_k(x)) \\ = \eta(G(x))(G_*(X_1(x)), \dots, G_*(X_k(x))), \end{aligned}$$

which is an exterior form in  $\Lambda^k(T_x^* V)$ . In the case of linear forms, this is the dual mapping to the differential  $D^1 G$ . We can compute directly from the definition that, for instance,

$$G^*(dy_i) \left( \frac{\partial}{\partial x_k} \right) = dy_i \left( G_* \left( \frac{\partial}{\partial x_k} \right) \right) = \frac{\partial g_i}{\partial x_k},$$

and so

$$(1) \quad G^*(dy_i) = \frac{\partial g_i}{\partial x_1} dx_1 + \dots + \frac{\partial g_i}{\partial x_m} dx_m,$$

which extends to the linear combinations of all  $dy_i$  over functions.

Another immediate consequence of the definition is the formula for pullbacks of arbitrary  $k$ -forms by composing two diffeomorphisms:

$$(2) \quad (G \circ F)^* \alpha = F^*(G^* \alpha).$$

Indeed, as a mapping on  $k$ -tuples of vectors,

$$\begin{aligned} (G \circ F)^* \alpha &= \alpha \circ ((D^1 G \circ D^1 F) \times \dots \times (D^1 G \circ D^1 F)) \\ &= G^*(\alpha) \circ (D^1 F \times \dots \times D^1 F) = F^* \circ G^* \alpha \end{aligned}$$

as expected.

**9.1.3. Wedge product of exterior forms.** Given a  $k$ -form



$\alpha \in \Lambda^k \mathbb{R}^{n*}$  and an  $\ell$ -form  $\beta \in \Lambda^\ell \mathbb{R}^{n*}$ , we can create a  $(k + \ell)$ -form  $\alpha \wedge \beta$  by all possible permutations  $\sigma$  of the arguments.

We just have to alternate the arguments in all possible orders and take the right sign each time:

$$\begin{aligned} (\alpha \wedge \beta)(X_1, \dots, X_{k+\ell}) = \\ \frac{1}{k! \ell!} \sum_{\sigma \in \Sigma_{k+\ell}} \text{sgn}(\sigma) \alpha(X_{\sigma(1)}, \dots, X_{\sigma(k)}) \beta(X_{\sigma(k+1)}, \dots, X_{\sigma(k+\ell)}). \end{aligned}$$

where  $c$  is the positively oriented circle  $x^2 + y^2 = 9$ .  $\circlearrowleft$

**9.B.8.** Compute the integral

$$\int_c \left( \frac{1}{x} + 2xy - \frac{y^3}{3} \right) dx + \left( \frac{1}{y} + x^2 + \frac{x^3}{3} \right) dy,$$

where  $c$  is the positively oriented boundary of the set  $D = \{(x, y) \in \mathbb{R}^2 : 4 \leq x^2 + y^2 \leq 9, \frac{x}{\sqrt{3}} \leq y \leq \sqrt{3}x\}$ .  $\circlearrowleft$

**9.B.9. Remark.** An important corollary of *Green's theorem* is the formula for computing the area  $D$  that is bounded by a curve  $c$ .

$$m(D) = \frac{1}{2} \int_c -y \, dx + x \, dy.$$

**9.B.10.** Compute the area given by the ellipse  $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ .

**Solution.** Using the formula 9.B.9 and the transformation  $x = a \cos t$ ,  $y = b \sin t$ , we get for  $t \in [0, 2\pi]$  that

$$\begin{aligned} m(D) &= \frac{1}{2} \int_c -y \, dx + x \, dy = \\ &= \frac{1}{2} \int_0^{2\pi} a \cos t \cdot b \cos t \, dt - \frac{1}{2} \int_0^{2\pi} b \sin t \cdot (-a \sin t) \, dt = \\ &= \frac{1}{2} ab \int_0^{2\pi} \cos^2 t \, dt + \frac{1}{2} ab \int_0^{2\pi} \sin^2 t \, dt = \\ &= \frac{1}{2} ab \int_0^{2\pi} \cos^2 t + \sin^2 t \, dt = \frac{1}{2} ab 2\pi = \pi ab, \end{aligned}$$

which is indeed the well-known formula for the area of an ellipse with semi-axes  $a$  and  $b$ .  $\square$

**9.B.11.** Find the area bounded by the cycloid which is given parametrically as  $\psi(t) = [a(t - \sin t); a(1 - \cos t)]$ , for  $a \geq 0$ ,  $t \in (0, 2\pi)$ , and the  $x$ -axis.

**Solution.** Let the curves that bound the area be denoted by  $c_1$  and  $c_2$ . As for the area, we get

$$m(D) = \frac{1}{2} \int_{c_1} -y \, dx + x \, dy + \frac{1}{2} \int_{c_2} -y \, dx + x \, dy.$$

Now, we will compute the mentioned integrals step by step. The parametric equation of the curve  $c_1$  (a segment of the  $x$ -axis) is  $(t; 0); t \in [0; 2a\pi]$ , so we obtain for the first integral that

$$\frac{1}{2} \int_{c_1} -y \, dx + x \, dy = \frac{1}{2} \int_0^{2a\pi} 0 \cdot 1 \, dt + \int_0^{2a\pi} t \cdot 0 \, dt = 0.$$

The parametric equation of the curve  $c_2$  is  $\psi(t) \in (a(t - \sin t), a(1 - \cos t)); t \in [2\pi; 0]$ .

The formula for the area expects a positively oriented curve,

It is clear from the definition that  $\alpha \wedge \beta$  is indeed a  $(k + \ell)$ -form. In the simplest case of 1-forms, the definition says that

$$(\alpha \wedge \beta)(X, Y) = \alpha(X)\beta(Y) - \alpha(Y)\beta(X).$$

In the case of a 1-form  $\alpha$  and a  $k$ -form  $\beta$ , we get

$$\begin{aligned} (\alpha \wedge \beta)(X_0, X_1, \dots, X_k) = \\ \sum_{j=0}^k (-1)^j \alpha(X_j) \beta(X_0, \dots, \hat{X}_j, \dots, X_k), \end{aligned}$$

where the hat indicates omission of the corresponding argument. The wedge product of finitely many forms is defined analogously (either directly by a similar formula, or we can notice that the above wedge product of forms is an associative operation – think this out by yourselves!).

Next, remind the generators  $\frac{\partial}{\partial x_i}$  of all vector fields in  $\mathcal{X}(\mathbb{R}^n)$ , as well as the generators  $dx_i$  of all linear exterior forms in  $\Omega^1(\mathbb{R}^n)$ . Their wedge products

$$\varepsilon_{i_1 \dots i_k} = dx_{i_1} \wedge \dots \wedge dx_{i_k}$$

with  $k$ -tuples of indices  $i_1 < i_2 < \dots < i_k$  generate the whole space  $\Omega^k(\mathbb{R}^n)$  by linear combinations with functions standing for coefficients. Indeed, interchanging a pair of adjacent forms in the product merely changes the sign, so the whole expression is identically zero if an index appears twice. Therefore, every  $k$ -form  $\alpha$  is given uniquely by functions  $\alpha_{i_1 \dots i_k}(x)$  in the expression

$$\alpha(x) = \sum_{i_1 < \dots < i_k} \alpha_{i_1 \dots i_k}(x) dx_{i_1} \wedge \dots \wedge dx_{i_k}.$$

Consequently, the vector spaces  $\Lambda^k(T^*U)$  are the trivial zero spaces if  $k > \dim U$ . Thus,  $\Omega^k(U)$  contains only the trivial zero form in this case.

Another straightforward consequence of the definition is that the pullback of the wedge product by a smooth mapping  $G : V \rightarrow U$  satisfies

$$G^*(\alpha \wedge \beta) = G^*\alpha \wedge G^*\beta.$$

We should also notice that 0-forms  $\Omega^0(\mathbb{R}^n)$  are just smooth functions on  $\mathbb{R}^n$ . The wedge product of a 0-form  $f$  and a  $k$ -form  $\alpha$  is just the multiple of the form  $\alpha$  by the function  $f$ . Similarly, the top degree forms in  $\Omega^n(U)$  are all generated by the single generator  $\varepsilon_{12\dots n}$ , since there is just one possibility of  $n$  different choices among  $n$  coordinates, up to the ordering. This means that actually the  $n$ -forms  $\omega$  are identified with functions via the formula

$$\omega(x) = f(x) dx_1 \wedge \dots \wedge dx_n.$$

At the same time, while the pullback on the functions  $f \in \Omega^0(U)$  by a transformation  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $y = F(x)$ , is trivial, i.e.  $F^*f(x) = f(y) = f \circ F(x)$ , a straightforward computation reveals

$$(1) \quad F^*\omega(x) = \det(D^1F)(x) f(F(x)) dx_1 \wedge \dots \wedge dx_n$$

for all  $\omega = f dy_1 \wedge \dots \wedge dy_n$ .

which means for the considered parametric equation that we are moving against the parametrization direction, i. e. from the upper bound to the lower one.

We thus get for the area of the cycloid that

$$\begin{aligned} \frac{1}{2} \int_{c_2} -y \, dx + x \, dy &= \frac{1}{2} \int_{2\pi}^0 a(t - \sin t) \cdot a(\sin t) \, dt - \\ &- \frac{1}{2} \int_{2\pi}^0 a(1 - \cos t) \cdot a(1 - \cos t) \, dt = \\ &= \frac{1}{2} a^2 \int_0^{2\pi} t \sin t - \sin^2 t - 1 + 2 \cos t - \cos^2 t \, dt = \\ &= \frac{1}{2} a^2 \int_{2\pi}^0 t \sin t + 2 \cos t - 2 \, dt = \\ &= \frac{1}{2} a^2 [-t \cos t - \sin t + 2 \cos t - 2]_{2\pi}^0 = 3\pi a^2. \end{aligned}$$

**9.B.12.** Compute  $I = \iint_S x^3 \, dy \, dz + y^3 \, dx \, dz + z^3 \, dx \, dy$ , where  $S$  is given by the sphere  $x^2 + y^2 + z^2 = 1$ .

**Solution.** It is advantageous to work in spherical coordinates

$$\begin{aligned} x &= \rho \sin \varphi \cos \psi & \rho &= [0, 1], \\ y &= \rho \sin \varphi \sin \psi & \varphi &= [0, \pi], \\ z &= \rho \cos \varphi & \psi &= [0, 2\pi]. \end{aligned}$$

The Jacobian of this transformation is  $-\rho^2 \sin \varphi$ .

The given integral is then equal to

$$\begin{aligned} I &= \iiint_S x^3 \, dy \, dz + y^3 \, dx \, dz + z^3 \, dx \, dy = \\ &= \iiint_V 3x^2 + 3y^2 + 3z^2 \, dx \, dy \, dz = \\ &= 3 \int_0^1 \int_0^{2\pi} \int_0^\pi \rho^2 \sin \varphi (\rho^2 \sin^2 \varphi \cos^2 \psi + \rho^2 \sin^2 \varphi \sin^2 \psi + \\ &\quad + \rho^2 \cos^2 \varphi) \, d\rho \, d\varphi \, d\psi = \\ &= 3 \int_0^1 \int_0^{2\pi} \int_0^\pi \rho^4 \sin \varphi (\sin^2 \varphi (\cos^2 \psi + \sin^2 \psi) + \\ &\quad + \cos^2 \varphi) \, d\rho \, d\varphi \, d\psi = \\ &= 3 \int_0^1 \int_0^{2\pi} \int_0^\pi \rho^4 \sin \varphi \, d\rho \, d\varphi \, d\psi = 3 \cdot \left[ \frac{\rho^5}{5} \right]_0^1 [\psi]_0^{2\pi} [\cos \varphi]_0^\pi = \\ &= 3 \cdot \frac{1}{5} \cdot 2\pi \cdot [-1 - 1] = -\frac{12}{5}\pi. \end{aligned}$$

**9.1.4. Integration of exterior forms on  $\mathbb{R}^n$ .** Once we fix coordinates  $(x_1, \dots, x_n)$  on  $\mathbb{R}^n$  (e.g. the standard ones), there is the bijection between functions  $f$  and top degree forms  $\omega(x) = f(x) dx_1 \wedge \dots \wedge dx_n$ . This can be interpreted as defining the scale with which the standard volume in  $\mathbb{R}^n$  is to be taken pointwise due to the function  $f$ .



Notice, that changing the coordinates via a transformation  $F$  will rescale this understanding of the forms exactly as in the formula for coordinate substitution in the Riemann integral. We should view this observation as a new interpretation of the integrands in our earlier procedure of integration of functions  $f$  on Riemann measurable open subsets  $U \subset \mathbb{R}^n$ , independent of any coordinate choice.

Let us check this interpretation in more detail. First, we define the  $n$ -form  $\omega_{\mathbb{R}^n}$ , giving the standard  $n$ -dimensional volume of parallelograms, i.e. in the standard coordinates we obtain

$$\omega_{\mathbb{R}^n} = dx_1 \wedge \dots \wedge dx_n.$$

If we want to integrate a function  $f(x)$  “in the new way”, we consider the form  $\omega = f\omega_{\mathbb{R}^n}$  instead, i.e.  $\omega = f(x) dx_1 \wedge \dots \wedge dx_n$ . We define the integral of the form  $\omega$  as

$$\int_U \omega = \int_U f(x) dx_1 \wedge \dots \wedge dx_n = \int_U f(x) \, dx_1 \cdots dx_n,$$

where the Riemann integral of a function is considered on the right-hand side.

Let us point out, that the  $n$ -form  $\omega$  on the left-hand side is well defined, independently of any choice of coordinates. If we want to express the form  $\omega$  in different coordinates using a diffeomorphism  $G : V \rightarrow U$ ,  $G = (g_1, \dots, g_n)$ , it means we will evaluate  $\omega$  at a point  $G(y) = x$  at the values of the vectors  $G_*(X_1), \dots, G_*(X_n)$ . However, this means we will integrate the form  $G^*\omega$  in coordinates  $(y_1, \dots, y_n)$ , and we already saw in the previous paragraph, cf. 9.1.3(1) that

$$(G^*\omega)(y) = f(G(y)) \det(D^1G(y)) dy_1 \wedge \dots \wedge dy_n.$$

Substituting into our interpretation of the integral, we get

$$\int_V G^*(f\omega_{\mathbb{R}^n}) = \int_{G^{-1}(U)} f(G(y)) \det(D^1G(y)) dy_1 \cdots dy_n,$$

which is, by the theorem 8.2.8 on the coordinate substitution in the integral, the same value as  $\int_U f\omega_{\mathbb{R}^n}$  if the determinant of the Jacobian matrix is positive, and the same value up to the sign if it is negative.

Our new interpretation thus provides the geometrical meaning for the integral of an  $n$ -form on  $\mathbb{R}^n$ , supposing the corresponding Riemann integral exists in some (and hence any) coordinates. This integration takes into account the orientation of the area we are integrating over. We shall come back to this point in a moment.

**9.1.5. Integrals along curves.** Our next goal is to integrate objects over domains which are similar to curves or surfaces in  $\mathbb{R}^3$ . Let us first shape our mind on the simplest case of the lowest



□ dimension, i.e. the curves in  $\mathbb{R}^n$ .

**9.B.13. The vector form of the Gauss–Ostrogradsky theorem.** The divergence of a vector field  $F(x, y, z) = f(x, y, z) \frac{\partial}{\partial x} + g(x, y, z) \frac{\partial}{\partial y} + h(x, y, z) \frac{\partial}{\partial z}$  is defined as  $\operatorname{div} X := f_x + g_y + h_z$ . Then, the Gauss–Ostrogradsky theorem can be formulated as follows:

$$\iiint_V \operatorname{div} \vec{F}(x, y, z) \, dx \, dy \, dz = \iint_S \vec{F}(x, y, z) \cdot \vec{n}(x, y, z) \, dS$$

where  $\vec{n}(x, y, z)$  is the outer unit normal to the surface  $S$  at the point  $[x, y, z] \in S$  ( $S$  is the boundary of the normal domain  $V$ ).

**9.B.14.** Find the flow of the vector field given by the function  $F = (xy^2, yz, x^2z)$  over the cylinder  $x^2 + y^2 = 4, z = 1, z = 3$ .

**Solution.** First of all, we compute the divergence of the vector field:

$$\operatorname{div} F = \nabla \cdot F = \left( \frac{\partial(xy^2)}{\partial x} + \frac{\partial(yz)}{\partial y} + \frac{\partial(x^2z)}{\partial z} \right) = y^2 + z + x^2.$$

Therefore, the flow  $T$  of the vector field is equal to

$$\begin{aligned} \iiint_V y^2 + z + x^2 \, dx \, dy \, dz &= \\ &= \int_0^2 \int_0^{2\pi} \int_1^3 \rho \cdot (\rho^2 \sin^2 \varphi + z + \rho^2 \cos^2 \varphi) \, d\rho \, d\varphi \, dz = \\ &= \int_0^2 \int_0^{2\pi} \int_1^3 \rho \cdot (\rho^2 (\sin^2 \varphi + \cos^2 \varphi) + z) \, d\rho \, d\varphi \, dz = \\ &= \int_0^2 \int_0^{2\pi} \int_1^3 \rho \cdot (\rho^2 (\sin^2 \varphi + \cos^2 \varphi) + z) \, d\rho \, d\varphi \, dz = \\ &= \int_0^2 \int_0^{2\pi} \int_1^3 \rho^3 + \rho z \, d\rho \, d\varphi \, dz = \\ &= 2\pi \int_0^2 \int_1^3 \rho^3 + \rho z \, d\rho \, dz = 2\pi \int_1^3 \left[ \frac{\rho^4}{4} + \frac{\rho^2}{2} z \right]_0^3 dz = \\ &= 2\pi \int_1^3 4 + 2z \, dz = 2\pi [4z + z^2]_1^3 = 2\pi [12 + 9 - 4 - 1] = 22\pi. \end{aligned}$$

**9.B.15.** Find the flow of the vector field given by the function  $F = (y, x, z^2)$ , over the sphere  $x^2 + y^2 + z^2 = 4$ .

**Solution.** The divergence of the given vector field is:

$$\operatorname{div} F = \nabla \cdot F = \left( \frac{\partial y}{\partial x} + \frac{\partial x}{\partial y} + \frac{\partial z^2}{\partial z} \right) = 2z.$$

Recall the calculation of the length of a curve in  $\mathbb{R}^n$  by univariate integrals, which was discussed in paragraph 6.1.10 on page 518. The curve was parametrized as a mapping  $c(t) : \mathbb{R} \rightarrow \mathbb{R}^n$ , and the size of the tangent vector  $\|c'(t)\|$  was expressed in the Euclidean vector space. This procedure was given by the universal relation for an arbitrary tangent vector, i.e., we actually found the function  $\rho : \mathbb{R}^n \rightarrow \mathbb{R}$  which gave the true size when evaluated at  $c'(t)$ . This mapping satisfied  $\rho(av) = |a|\rho(v)$  since we ignored the orientation of the curve given by our parametrization. If we wanted a signed length, respecting the orientation, then our mapping  $\rho$  would be linear on every one-dimensional subspace  $L \subset \mathbb{R}^n$ . Of course we could have multiplied the Euclidean size by a positive function and integrate this quantity.

In view of our geometric approach to integration, we should rather integrate linear forms along curves, while the size of vectors is given by a quadratic form, rather than a linear one. However, in dimension one, we take the square root of the values of the (positive definite) quadratic form, in order to get a linear form (up to sign) which is just the size of the vectors.

Let us proceed in a much similar way dealing with linear differential forms  $\eta$  on  $\mathbb{R}^n$ . The simplest ones are the differentials  $df$  of functions  $f$  on  $\mathbb{R}^n$ .

In order to motivate our development, let us consider the following task. Imagine, we are cycling along a path  $c(t)$  in  $\mathbb{R}^2$ , the function  $f$  is the altitude of the terrain. If we want to compute the total gain of altitude along the path  $c(t)$ , we should “integrate” the immediate infinitesimal gains, which should be the derivatives of  $f$  in the directions of the tangent vectors to the path, i.e.  $df(c'(t))$ .

Thus, let us consider a differentiable curve  $c(t)$  in  $\mathbb{R}^n$ ,  $t \in [a, b]$ , write  $M$  for the image  $c([a, b])$ , and assume that a differentiable function  $f$  is defined on a neighborhood of  $M$ . The differential of this function gives for every tangent vector the increment of the function in the given direction. It is expressed by the differential of the composite mapping  $f \circ c$

$$d(f \circ c)(t) = \frac{\partial f}{\partial x_1}(c(t))c'_1(t) + \dots + \frac{\partial f}{\partial x_n}(c(t))c'_n(t).$$

We can thus try to define the value of the integral in the following way

$$\int_M df = \int_a^b \left( \frac{\partial f}{\partial x_1}(c(t))c'_1(t) + \dots + \frac{\partial f}{\partial x_n}(c(t))c'_n(t) \right) dt,$$

and we immediately verify that the change of the parametrization of the curve has no effect upon the value. Indeed, writing  $c(t) = c(\psi(s))$ ,  $a = \psi(\tilde{a})$ ,  $b = \psi(\tilde{b})$ , our procedure yields

$$\begin{aligned} &\int_{\tilde{a}}^{\tilde{b}} \left( \frac{\partial f}{\partial x_1}(c(\psi(s)))c'_1(\psi(s)) + \dots \right. \\ &\quad \left. + \frac{\partial f}{\partial x_n}(c(\psi(s)))c'_n(\psi(s)) \right) \frac{d\psi}{ds} ds, \end{aligned}$$

and the theorem about coordinate transformations for univariate integrals gives just the same value if we have  $\frac{d\psi}{ds} > 0$ , i.e.,

Thus, the wanted flow equals

$$\begin{aligned} \iiint_V 2z \, dx \, dy \, dz &= \int_0^2 \int_0^\pi \int_0^{2\pi} \rho^2 \sin \varphi \cdot 2\rho \cos \varphi \, d\rho \, d\varphi \, d\psi \\ &= 2 \int_0^2 \rho^3 \, d\rho \int_0^{2\pi} d\psi \int_0^\pi \sin \varphi \cos \varphi \, d\varphi = \\ &= 2 \left[ \frac{\rho^4}{4} \right]_0^2 \cdot [\psi]_0^{2\pi} \cdot \left[ \frac{\sin^2 \varphi}{2} \right]_0^\pi = \\ &= 2 \cdot \frac{16}{4} \cdot 2\pi \cdot 0 = 0. \end{aligned}$$

### C. Equation of heat conduction

**9.C.1.** Find the solution to the so-called equation of heat conduction (equation of diffusion)

$$u_t(x, t) = a^2 u_{xx}(x, t), \quad x \in \mathbb{R}, \quad t > 0$$

satisfying the initial condition  $\lim_{t \rightarrow 0^+} u(x, t) = f(x)$ .

Notes: The symbol  $u_t = \frac{\partial u}{\partial t}$  stands for the partial derivative of the  $u$  with respect to  $t$  (i. e., differentiating with respect to  $t$  and considering  $x$  to be constant), and similarly,  $u_{xx} = \frac{\partial^2 u}{\partial x^2}$  denotes the second partial derivative with respect to  $x$  (i. e., twice differentiating with respect to  $x$  while considering  $t$  to be constant). The physical interpretation of this problem is as follows: We are trying to determine the temperature  $u(x, t)$  in an thermally isolated and homogeneous bar of infinite length (the range of the variable  $x$ ) if the initial temperature of the bar is given as the function  $f$ . The section of the bar is constant and the heat can spread in it by conduction only. The coefficient  $a^2$  then equals the quotient  $\frac{\alpha}{c\rho}$ , where  $\alpha$  is the coefficient of thermal conductivity,  $c$  is the specific heat and  $\rho$  is the density. In particular, we assume that  $a^2 > 0$ .

**Solution.** We apply the Fourier transform to the equation, with respect to variable  $x$ . We have

$$\begin{aligned} \mathcal{F}(u_t)(\omega, t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u_t(x, t) e^{-i\omega x} \, dx = \\ &= \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x, t) e^{-i\omega x} \, dx \right)', \end{aligned}$$

where differentiated with respect to  $t$ , i. e.,

$$\mathcal{F}(u_t)(\omega, t) = (\mathcal{F}(u)(\omega, t))' = (\mathcal{F}(u))_t(\omega, t).$$

At the same time, we know that

$$\begin{aligned} \mathcal{F}(a^2 u_{xx})(\omega, t) &= a^2 \mathcal{F}(u_{xx})(\omega, t) = \\ &= -a^2 \omega^2 \mathcal{F}(u)(\omega, t). \end{aligned}$$

Denoting  $y(\omega, t) = \mathcal{F}(u)(\omega, t)$ , we get to the equation

$$y_t = -a^2 \omega^2 y.$$

We already solved a similar differential equation when we were calculating Fourier transforms, so it is now easy for us to determine all of its solutions

$$y(\omega, t) = K(\omega) e^{-a^2 \omega^2 t}, \quad K(\omega) \in \mathbb{R}.$$

if we keep the orientation of the curve, and the same value up to sign if the derivative of the transformation is negative.

If we extend the same definition to an arbitrary linear form  $\eta = \eta_1 dx_1 + \dots + \eta_n dx_n$  we arrive at the same formulae with  $\eta_i$  replacing the derivatives  $\frac{\partial f}{\partial x_i}$ ,

$$\int_M \eta = \int_a^b (\eta_1(c(t))c'_1(t) + \dots + \eta_n(c(t))c'_n(t)) dt,$$

again independent of the parametrization of the curve  $c$  as above.

In the above example with  $n = 2$ ,  $f$  was the altitude of the terrain, and the integral of  $df$  along the path modelled the total gain of elevation. Thus, we should expect that the total gain along the path should depend on the values  $c(a)$  and  $c(b)$  only, while different curves with the same boundary points would produce different integrals of  $\eta$  for a general 1-form  $\eta$ . This will indeed be the special claim of the Stokes theorem below.

Before we treat the higher dimensional analogs, we shall look at more abstract approach to suitable subsets in  $\mathbb{R}^n$  and the role of coordinates on them.

**9.1.6. Manifolds.** The straightforward generalizations of parameterized curves  $c(t) : \mathbb{R} \rightarrow \mathbb{R}^n$  are the differentiable mappings  $\varphi : V \subset \mathbb{R}^k \rightarrow \mathbb{R}^n$ ,  $k \leq n$ ,



with injective differential  $d\varphi(u)$  at every point of its open domain  $V$ . Such mappings are called *immersions*.

With the curves, we did not care about their self-intersections etc. Now, for technical reasons, we shall be more demanding.

#### MANIFOLDS IN $\mathbb{R}^n$

A subset  $M \subset \mathbb{R}^n$  is called a *manifold of dimension  $k$*  if every point  $x \in M$  has a neighborhood  $U \subset \mathbb{R}^n$  which is the image of a diffeomorphism  $\tilde{\varphi} : V \times \tilde{V} \rightarrow U$ ,  $V \subset \mathbb{R}^k$ ,  $\tilde{V} \subset \mathbb{R}^{n-k}$ , such that

- the restriction  $\varphi = \tilde{\varphi}|_V : V \rightarrow M$  is an immersion,
- $\tilde{\varphi}^{-1}(M) = V \times \{0\} \subset \mathbb{R}^n$ .

The manifolds  $M$  are carrying the topology inherited from  $\mathbb{R}^n$ .

./img/0214\_eng.png

It remains to determine  $K(\omega)$ . The transformation of the initial condition gives

$$\mathcal{F}(f)(\omega) = \lim_{t \rightarrow 0^+} \mathcal{F}(u)(\omega, t) = \lim_{t \rightarrow 0^+} y(\omega, t) = K(\omega) e^0 = K(\omega),$$

hence

$$y(\omega, t) = \mathcal{F}(f)(\omega) e^{-a^2 \omega^2 t}, \quad K(\omega) \in \mathbb{R}.$$

Now, using the inverse Fourier transform, we can return to the original differential equation with solution

$$\begin{aligned} u(x, t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y(\omega, t) e^{i\omega x} d\omega = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathcal{F}(f)(\omega) e^{-a^2 \omega^2 t} e^{i\omega x} d\omega = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(s) e^{-i\omega s} ds \right) e^{-a^2 \omega^2 t} e^{i\omega x} d\omega = \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(s) \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-a^2 \omega^2 t} e^{-i\omega(s-x)} d\omega \right) ds. \end{aligned}$$

Computing the Fourier transform  $\mathcal{F}(f)$  of the function  $f(t) = e^{-at^2}$  for  $a > 0$ , we have obtained (while relabeling the variables)

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-cp^2} e^{-irp} dp = \frac{1}{\sqrt{2c}} e^{-\frac{r^2}{4c}}, \quad c > 0.$$

According to this formula (consider  $c = a^2 t > 0$ ,  $p = \omega$ ,  $r = s - x$ ), we have

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-a^2 \omega^2 t} e^{-i\omega(s-x)} d\omega = \frac{1}{\sqrt{2a^2 t}} e^{-\frac{(s-x)^2}{4a^2 t}},$$

Therefore,

$$u(x, t) = \frac{1}{2a\sqrt{\pi t}} \int_{-\infty}^{\infty} f(s) e^{-\frac{(x-s)^2}{4a^2 t}} ds.$$

□

### D. Partial Differential Equations

**9.D.1.** Find general solution and solution for given boundary conditions of homogenous linear equation

$$xu_x + yu_y = 0,$$

- (i)  $u(\cos \sigma, \sin \sigma) = 1$ ,
- (ii)  $u(\sigma, 1) = 1 + \sigma^2$ ,
- (iii)  $u(\sigma, 1 - \sigma) = 3\sigma$ ,
- (iv)  $u(\cos \sigma, \sin \sigma) = \sigma$ ,
- (v)  $u(\sigma, \sigma) = 1$ ,
- (vi)  $u(\sigma, \sigma) = \sigma$ .

**Solution.** Firstly let's look to our problem geometrically. The vector  $\nabla u = (u_x, u_y)$  is the *gradient* of unknown function  $u = u(x, y)$  (its direction is direction of the biggest growing of function  $u$ ). Equation

$$a(x, y)u_x + b(x, y)u_y = 0,$$

tell us, that scalar product of given vector field  $\vec{v} = (a(x, y), b(x, y))$  and  $\nabla u$  is zero, e.g.  $\vec{v}$  is orthogonal to  $\nabla u$  at each point, so  $\vec{v}$  has to be tangent to the contour lines (lines of constant value of  $u$ ), we will call them

This definition is illustrated by the picture above. Manifolds can be typically (at least locally) given implicitly as the level sets of differentiable mappings, see paragraph 8.1.24 and the discussion in 8.1.26.

The mapping  $\varphi$  from the definition is called the *local parametrization* or *local map* of the manifold  $M$ . The manifolds are a straightforward generalization of curves and surfaces in the plane  $\mathbb{R}^2$  or the space  $\mathbb{R}^3$ . We have excluded curves and surfaces which are self-intersecting and even those which are self-approaching.

For instance, we can surely imagine a curve representing the figure 8 parametrized with a mapping  $\varphi$  with everywhere-injective differential. However, we will be unable to satisfy the second property from the manifold definition in a neighborhood of the point where the two branches of the curve meet.

### TANGENT AND COTANGENT BUNDLES OF MANIFOLDS

The tangent bundle  $TM$  of the manifold  $M$  is the collection of vector subspaces  $T_x M \subset T_x \mathbb{R}^n$  which contain all vectors tangent to the curves in  $M$ . There is the footpoint projection  $p: TM \rightarrow M$ .

Similarly, the cotangent bundle  $T^*M$  of the manifold  $M$  is the collection of the dual spaces  $(T_x M)^*$ , together with the footpoint projection.



Clearly, every parametrization  $\varphi$  defines a diffeomorphism

$$\varphi_*: TV \rightarrow T(\varphi(V)) \subset TM, \quad \varphi_*(c'(t)) = \frac{d}{dt} \varphi(c(t)).$$

Due to the chain rule, this definition does not depend of the choice of the representing curve  $c(t)$ . We shall also write  $T\varphi$  for the mapping  $\varphi_*$ .

In particular, the local maps  $\varphi$  (extended to  $\tilde{\varphi}$ , as in the above definition) induce the local maps  $\varphi_*: TU = U \times \mathbb{R}^k \rightarrow TM \subset \mathbb{R}^n \times \mathbb{R}^n$  of the tangent bundle. Thus, the tangent bundle  $TM$  is again a manifold, which locally looks as  $U \times \mathbb{R}^k$  over sufficiently small open subsets  $U \subset M$ . But we shall see that  $TM$  might be quite different from  $M \times \mathbb{R}^k$  globally. Dealing with the cotangent bundle, we can use the dual mappings  $(T\varphi^{-1})^*$  on the individual fibers  $T_x^* M$  to obtain local parametrizations.

*characteristics.* Now it's easy to find general solution — it could be any function, which is constant on characteristics (integral curves of vector field  $\vec{v}$ ). Let's find them for our case by solving system ODE's:

$$\dot{x}(t) = \frac{dx}{dt}(t) = a(x, y) = x, \dot{y}(t) = \frac{dy}{dt}(t) = b(x, y) = y,$$

or by solving equation

$$y'(x) = \frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{b}{a}.$$

Solution yields to the curves

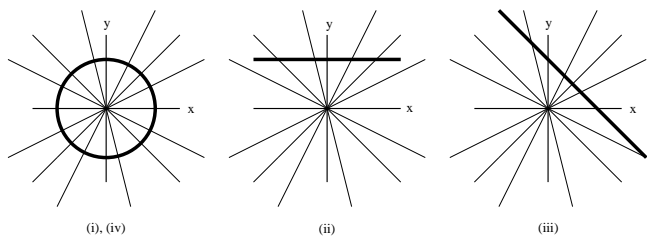
$$x = C_1 e^t, \quad y = C_2 e^t, \quad \text{or} \quad \frac{y}{x} = C$$

and characteristics are lines throu the origin (picture P.1). General solution can be written as

$$u(x, y) = \Phi\left(\frac{y}{x}\right), \quad \text{or} \quad u(x, y) = \Psi\left(\frac{x}{y}\right).$$

Look at the test:

$$u_x = \Phi'\left(-\frac{y}{x^2}\right), u_y = \Phi'\left(\frac{1}{x}\right) \implies xu_x + yu_y = 0.$$



On the picture we see characteristics together with given boundary lines. We need to choose solutions satisfying our boundary condition. The questions are: How many solutions we can find? Does the solution exists for any boundary condition?

- (i) Parametrizet boundary curve (circle) intersect any characteristics infinitely many times. But the value  $u(\cos \sigma, \sin \sigma) = 1$  is constant. Solution has to be constant on characteristics, therefore  $u(x, y) = 1$  is unique solution of the equation and boundary condition (i).
- (ii) Parametrized boundary curve (line)  $x = \varphi(\sigma) = \sigma, y = \psi(\sigma) = 1$  intersects almost all characteristic (except line  $y = 0$ ) just once. The value  $u = f(\sigma) = 1 + \sigma^2$  in point of intersection should be the same on the whole characteristics: Without lose of generality, take  $t = 0$  and from  $x = C_1 e^0 = \varphi(\sigma) = \sigma, y = C_2 e^0 = \psi(\sigma) = 1$  we get  $C_1 = \sigma, C_2 = 1$ . Then

$$x = \sigma e^t, \quad y = e^t \implies \sigma = \frac{x}{y}$$

and we have solution

$$u(x, y) = f(\sigma(x, y)) = 1 + [\sigma(x, y)]^2 = 1 + \frac{x^2}{y^2}.$$

This solution exists for  $y \neq 0$ . On the connected component of its domain, including boundary curve ( $\Omega =$

Notice that two differentiable immersions  $\varphi$  and  $\psi$  parametrizing the same open subset  $U \subset M$  provide the composition  $\psi^{-1} \circ \varphi$ . We view this as a coordinate change for  $U$  and we have just seen that coordinate changes on  $M$  induce coordinate changes on  $TM$ .

Further, if  $M$  and  $N$  are two manifolds and  $F : M \rightarrow N$  a mapping, we say that  $F$  is differentiable (up to order  $r$  or smooth or analytic), if the compositions  $\psi^{-1} \circ F \circ \varphi$  with two local parametrizations  $\varphi$  of  $M$  and  $\psi$  of  $N$  (of the same order of differentiability as we want to check) is differentiable (up to order  $r$  or smooth or analytic). Again, the chain rule property of differentiation shows that this definition does not depend on the particular choice of the parametrizations.

Each differentiable mapping  $F : M \rightarrow N$  defines the tangent mapping  $TF : TM \rightarrow TN$  between the tangent spaces, which clearly is differentiable of order one less than the assumed differentiability of  $F$ .

VECTOR FIELDS AND DIFFERENTIAL FORMS ON MANIFOLDS

Smooth vector fields  $X$  on a manifold  $M$  are smooth sections  $X : M \rightarrow TM$  of the footpoint projection  $p : TM \rightarrow M$ , i.e.,  $p \circ X = \text{id}_M$ .

Smooth  $k$ -forms  $\eta$  on a manifold  $M$  are sections  $M \rightarrow \Lambda^k(TM)^*$  such that the pullback of this form by any parametrization  $V \rightarrow M$  yields a smooth exterior  $k$ -form on  $V$ .

We write  $\mathcal{X}(M)$  for the space of smooth vector fields on  $M$ , while  $\Omega^k(M)$  stays for the space of all smooth exterior  $k$ -forms on  $M$ .

Notice that all our coordinate formulae for the vector fields, forms, pullbacks etc. on  $\mathbb{R}^m$  hold true in the more abstract setting of manifolds and their local parametrizations.<sup>1</sup>

9.1.7. Integration of exterior forms on manifolds. Now,



we are almost ready for the definition of the integral of  $k$ -forms on  $k$ -dimensional manifolds. For the sake of simplicity, we will examine smooth forms  $\omega$  with compact support only.

First, let us assume that we are given a  $k$ -dimensional manifold  $M \subset \mathbb{R}^n$  and one of its local parametrizations  $\varphi : V \subset \mathbb{R}^k \rightarrow U \subset M \subset \mathbb{R}^n$ . We consider the standard orientation on  $\mathbb{R}^k$  given by the standard basis (cf. 4.1.19 for the definition of the orientation of a vector space). The choice of the parametrization  $\varphi$  also fixes the *orientation of the manifold*  $U \subset M$ . This orientation will be the same for those choices of local parametrizations, which differ by diffeomorphisms with positive determinants of their Jacobi matrices. The orientation will be the other one in the case of negative determinants. The manifold  $M$  is called *orientable* if there

<sup>1</sup>Actually, instead of dealing with manifolds as subsets of  $\mathbb{R}^n$ , we might use the same concept of local parametrizations of a space  $M$  with differentiable transition functions  $\psi^{-1} \circ \varphi$ . We just need to know what are the "open subsets" in  $M$ , thus we could start at the level of topological spaces. On the other hand, there is the general result (the so called Whitney embedding theorem) that each such abstract  $n$ -dimensional manifold can be realized as embedded in the  $\mathbb{R}^{2n}$ , so we essentially do not loose any generality here.

$\{(x, y) \in \mathbb{R}^2 \mid y > 0\}$  this is unique continuous solution. There is no "natural" way, how to make its prolongation for  $y < 0$ .

- (iii) Parametrized boundary curve (line)  $x = \varphi(\sigma) = \sigma$ ,  $y = \psi(\sigma) = 1 - \sigma$  intersects almost all characteristic (except line  $y = -x$ ) just once. Again, the value  $u = f(\sigma) = 3\sigma$  in point of intersection should be the same on the whole characteristics: Take  $t = 0$  and from  $x = C_1 e^0 = \varphi(\sigma) = \sigma$ ,  $y = C_2 e^0 = \psi(\sigma) = 1$  we get  $C_1 = \sigma$ ,  $C_2 = 1 - \sigma$ . Then  $x = \sigma e^t$ ,  $y = (1 - \sigma)e^t$

$$\implies \frac{\sigma}{1 - \sigma} = \frac{x}{y}, \implies \sigma = \frac{x}{x + y}.$$

and we have the unique solution

$$u(x, y) = f(\sigma(x, y)) = 3\sigma(x, y) = \frac{3x}{x + y}.$$

This solution exists for  $y \neq -x$ . Let's make a test:

$$u_x = \frac{3y}{(x + y)^2}, \quad u_y = \frac{-3x}{(x + y)^2} \implies \implies xu_x + yu_y = 0,$$

$$u(\sigma, 1 - \sigma) = \frac{3\sigma}{\sigma + 1 - \sigma} = 3\sigma.$$

- (iv) Boundary curve intersect any characteristics infinitely many times, but the value at two point on the same characteristic is different. Globally this problem has no solution. We can solve it for example for  $\sigma \in (-\frac{\pi}{2}, \frac{\pi}{2})$ :  $t = 0$ ,  $x = C_1 e^0 = \cos \sigma$ ,  $y = C_2 e^0 = \sin \sigma$ , thus  $C_1 = \cos \sigma$ ,  $C_2 = \sin \sigma$  and

$$x = \cos \sigma e^t, \quad y = \sin \sigma e^t \implies \sigma = \arctg \frac{y}{x}.$$

We get the solution for  $x > 0$ :

$$u(x, y) = \sigma(x, y) = \arctg \frac{y}{x}.$$

Make a test.

- (v) The boundary curve  $y = x$  is one of the characteristics. It has no sense. Condition  $u(\sigma, \sigma) = 1$  gives infinitely many solutions. Any function  $u = \Phi(\frac{y}{x})$  or  $\Psi(\frac{x}{y})$  giving value 1 for  $y = x$  is solution of this problem (for example  $u = \frac{y}{x}$ ,  $u = \frac{y^2}{x^2}$ ,  $u = 1, \dots$ ).
- (vi) The boundary curve  $y = x$  is again one of the characteristics. It has no sense. There is no solution, because condition  $u(\sigma, \sigma) = \sigma$  doesn't give the same value on characteristic  $y = x$ .

Finally it seems (for our equation): If the boundary curve intersects each characteristics at exactly one point, there exists unique continuous solution (on some open set including boundary curve) satisfying boundary condition. Generally is situation more complicated. Requirement, that boundary curve has to intersect each characteristics at exactly one point is not necessary, nor sufficient).  $\square$

**9.D.2.** Find general solution of

$$yu_x + xu_y = 0$$

and solutions for given boundary condition

is a covering of the entire set  $M$  by local parametrizations  $\varphi$  such that their orientations coincide.

Therefore, we apparently have exactly two orientations on every connected orientable manifold. Fixing either of them, we thereby restrict the set of parametrizations to those compatible with this orientation. From now on, we will always proceed in this fashion, and we will talk about *oriented manifolds* only.

Next, let us fix a form  $\omega$  with compact support inside the image of one parametrization  $U \subset M$  of an oriented manifold  $M$ . The pullback form  $\varphi^*(\omega)$  is a smooth  $k$ -form on  $V \subset \mathbb{R}^k$  with compact support. The *integral of the form  $\omega$  on  $M$*  is defined in terms of the chosen parametrization which is compatible with the orientation as follows:

$$\int_M \omega = \int_{\mathbb{R}^k} \varphi^*(\omega).$$

If we choose a different compatible parametrization  $\tilde{\varphi} = \varphi \circ \psi$  where  $\psi$  is a diffeomorphism  $\psi : W \rightarrow V \subset \mathbb{R}^k$ , we can easily compute the result, following the same definition. Let us denote

$$\varphi^*(\omega)(u) = f(u) du_1 \wedge \dots \wedge du_k.$$

Invoking the relation 9.1.2(2) for the pullback of a form by a composite mapping, we get

$$\begin{aligned} \int_M \omega &= \int_{\mathbb{R}^k} \tilde{\varphi}^*(\omega) = \int_{\mathbb{R}^k} \psi^*(\varphi^*\omega) \\ &= \int_{\mathbb{R}^k} \psi^*(f du_1 \wedge \dots \wedge du_k) \\ &= \int_{\mathbb{R}^k} f(\psi(v)) \det(D^1\psi)(v) dv_1 \dots dv_k. \end{aligned}$$

This is again the same value as  $\int_{\mathbb{R}^k} \varphi^*\omega$ .

This proves the correctness of our definition of the integral  $\int_M \omega$  provided the integrated  $k$ -form has compact support lying in the image of a single parametrization.

However, typical manifolds  $M$  are given by implicit equations. For example,  $x^2 + y^2 + z^2 = 1$  defines the surface of the unit ball, i.e., the sphere  $S^2 \subset \mathbb{R}^3$ . If we want to integrate an exterior 2-form on  $S^2$ , we will have to use several parametrizations. Fortunately, our definition of the integral is additive with respect to disjoint unions of integration domains. Therefore, if we can write

$$M = U_1 \cup U_2 \cup \dots \cup U_m \cup B,$$

where  $U_i$  are pairwise disjoint images of parametrizations  $\varphi_i$ , and  $B$  is a set whose inverse image in any parametrization is a Riemann measurable set with measure zero, we can compute

$$\int_M \omega = \int_{U_1} \omega + \dots + \int_{U_m} \omega,$$

and we can easily verify that this value is independent of the choice of the sets  $U_i$  and the parametrizations (in particular, we need not be worried by the set  $B$  since the result of any integration on it is zero). For example, we can imagine splitting a sphere to the upper and lower hemispheres, leaving the equator  $B$  uncovered.



- (i)  $u(x, 0) = |x|$ ,
- (ii)  $u(0, y) = y^2$ ,
- (iii)  $u(\sigma, \sigma) = 2\sigma^2$ .

**9.D.3.** Find general solution of

$$yu_x - xu_y = 0$$

and solutions for given boundary condition

- (i)  $u(\sigma, \sigma) = 2\sigma^2$ ,
- (ii)  $u(0, y) = \frac{1}{1+y^2}$ ,
- (iii)  $u(x, 1) = x^4$ .

**9.D.4.** Find solution of

$$\sin x \sin y u_x + \cos x \cos y u_y = 0, u = \cos 2y \text{ for } x+y = \frac{\pi}{2}.$$

**9.D.5.** Look on general *quasilinear* first order equation:

$$a(x, y, u)u_x + b(x, y, u)u_y = f(x, y, u),$$

its special case is equation *linear in highest order terms* (in some texts, including our, is called *semilinear*):

$$a(x, y)u_x + b(x, y)u_y = f(x, y, u),$$

and again its special case is *linear* equation (generally nonhomogenous):

$$a(x, y)u_x + b(x, y)u_y = f(x, y).$$

In the following sections we show, how to solve each of this type of equation.

Find a solution of quasilinear equation for given boundary condition:

$$u_x - uu_y = -u, \quad u(0, y) = 2y.$$

**Solution.** Let's solve characteristic system:

$$\dot{x} = a = 1, \quad \dot{y} = b = -u, \quad \dot{u} = f = -u.$$

Characteristics are given parametrically:

$$x(t) = t + C_1, \quad y(t) = C_3 e^{-t} + C_2, \quad u(t) = C_3 e^{-t}.$$

If we want general solution of equation, it is given implicitly by

$$\Phi(ue^x, u - y) = 0,$$

where  $u - y = -C_2$ ,  $ue^x = C_3 e^{C_1}$  is some implicit description of characteristics.

From boundary condition for  $t = 0$  we get

$$x(0) = C_1 = 0, \quad y(0) = C_3 + C_2 = \sigma, \quad u(0) = C_3 = 2\sigma,$$

so  $C_2 = -\sigma$  and

$$x(t, \sigma) = t, \quad y(t, \sigma) = \sigma(2e^{-t} - 1) \implies$$

$$\sigma = \frac{y}{2e^{-x} - 1}, \quad e^{-t} = e^{-x}$$

When calculating in practice, we usually divide the entire manifold into several disjoint open areas with compact closures, and we integrate on each of them separately. However, this procedure still does not help if we stick with the strict assumption that the entire support of integrated form has to be inside of one parametrization. Thus, we will develop a global definition of the integral, which is more advantageous from the technical/theoretical point of view (although it usually does not help in computations directly).

**9.1.8. Partition of unity.** Consider a manifold  $M \subset \mathbb{R}^n$  and one of its covers by open images  $U_i$  of parametrizations  $\varphi_i$ . We can surely find a countable cover of each manifold  $M$  (it suffices to realize that we can do with parametrizations which map the origin to points with rational coordinates in  $\mathbb{R}^n$ ). Furthermore, we shall assume that any point in  $x \in M$  belongs to only finitely many sets  $U_i$ . Such a cover is called a *locally finite cover* by parametrizations  $\varphi_i$ .<sup>2</sup>



Now, recall the smooth variants of indicator functions from paragraph 6.1.10. For every pair of positive numbers  $\varepsilon < r$ , we constructed a function  $f_{\varepsilon, r}(t)$  of one real variable  $t$  such that  $f_{\varepsilon, r}(t) = 1$  for  $|t| < r - \varepsilon$ , while  $f_{\varepsilon, r}(t) = 0$  for  $|t| > r + \varepsilon$ , and  $0 \leq f_{\varepsilon, r}(t) \leq 1$  everywhere. At the same time, we had  $f(t) \neq 0$  if and only if  $|t| < r + \varepsilon$ .

Next, if we define

$$\chi_{r, \varepsilon, x_0}(x) = f_{\varepsilon, r}(|x - x_0|),$$

then we get a smooth function which takes the value 1 inside the ball  $B_{r-\varepsilon}(x_0)$ , with support exactly  $B_{r+\varepsilon}(x_0)$ , and with values between 0 and 1 everywhere.

mozna obr. char.fce

**Lemma** (Whitney's theorem). *Every closed set  $K \subset \mathbb{R}^n$  is the set of all zero points of some smooth non-negative function.*

**PROOF.** The idea of the proof is quite simple. If  $K = \mathbb{R}^n$ , the zero function fulfills the conditions, so we can further assume that  $K \neq \mathbb{R}^n$ .

The open set  $U = \mathbb{R}^n \setminus K$  can be expressed as the union of (at most) countably many open balls  $B_{r_i}(x_i)$ , and for each of them, we choose a smooth non-negative function  $f_i$  on  $\mathbb{R}^n$  whose support is just  $\overline{B_{r_i}(x_i)}$ , see the function  $\chi_{r, \varepsilon, x_0}$  above. Now, we add up all these functions into an infinite series

$$f(x) = \sum_{k=1}^{\infty} a_k f_k(x),$$

where the positive coefficients  $a_k$  are selected so small that this series would converge to a smooth function  $f(x)$ .

To this purpose, it suffices to choose  $a_k$  so that all partial derivatives of all functions  $a_k f_k(x)$  up to order  $k$  (inclusive) would be bounded from above by  $2^{-k}$ . Then, the series  $\sum_k a_k f_k$  is bounded from above by the series  $\sum_k 2^{-k}$ , hence by Weierstrass criterion, it converges uniformly on the entire

<sup>2</sup>This property is called *paracompactness* and, actually, each metric space is paracompact. Thus in particular all our manifolds enjoy this property too. But we do not want to go into details of the proof.

and

$$u(t, \sigma) = 2\sigma e^{-t} \implies u(x, y) = \frac{2y}{2 - e^x}.$$

This is unique continuous solution on connected component of its domain including boundary curve (for  $x < \ln 2$ ). Make a test.  $\square$

**9.D.6.** Find a solution of equation for boundary conditions (i)–(iii):

$$yu_x + xu_y = 2u$$

- (i)  $u(x, 0) = 1,$
- (ii)  $u(x, 0) = x^2,$
- (iii)  $u(0, y) = \sqrt{1 + y^2}.$

**Solution.** This is semilinear equation, we will solve it like quasilinear (in the next section we show another method working for semilinear equation):

$$\dot{x} = y, \dot{y} = x, \dot{u} = 2u.$$

Characteristics are given parametrically

$$x(t) = C_1 e^t + C_2 e^{-t}, y(t) = C_1 e^t - C_2 e^{-t}, u(t) = C_3 e^{2t}.$$

- (i) Without lose of generality put  $t = 0$  for the point of intersection of characteristic curve and boundary curve  $x = \varphi(\sigma) = \sigma, y = \psi(\sigma) = 0, u = f(\sigma) = 1$  (boundary curve intersects characteristics for  $x > y$ ),

$$x(0) = C_1 + C_2 = \sigma, y(0) = C_1 - C_2 = 0.$$

Then  $C_1 = C_2 = \frac{\sigma}{2}$  and from

$$x = \frac{\sigma}{2} (e^t + e^{-t}) = \sigma \cosh t, y = \frac{\sigma}{2} (e^t - e^{-t}) = \sigma \sinh t$$

we get  $x^2 - y^2 = \sigma^2, x + y = \sigma e^t,$

$$\sigma = \sqrt{x^2 - y^2}, \quad e^t = \frac{x + y}{\sqrt{x^2 - y^2}}.$$

$$u(0) = C_3 = f(\sigma) = 1,$$

$$u(x, y) = u(\sigma(x, y), t(x, y)) = e^{2t} = \left[ \frac{x + y}{\sqrt{x^2 - y^2}} \right]^2$$

Finally

$$u(x, y) = \frac{x + y}{x - y}, \text{ for } x > y.$$

Make a test.

- (ii) Boundary curve is the same,  $u(0) = C_3 = f(\sigma) = \sigma^2,$  solution is

$$u(x, y) = u(\sigma(x, y), t(x, y)) = \sigma^2 e^{2t} = (x + y)^2, \text{ for } x > y.$$

- (iii) Again put  $t = 0$  for the point of intersection of characteristic curve and boundary curve  $x = \varphi(\sigma) = 0, y = \psi(\sigma) = \sigma, u = f(\sigma) = \sqrt{1 + \sigma^2}$  (boundary curve intersects characteristics for  $y > x$ ),

$$x(0) = C_1 + C_2 = 0, y(0) = C_1 - C_2 = \sigma.$$

Then  $C_1 = -C_2 = \frac{\sigma}{2}$  and from

$$x = \frac{\sigma}{2} (e^t - e^{-t}) = \sigma \sinh t, y = \frac{\sigma}{2} (e^t + e^{-t}) = \sigma \cosh t$$

$\mathbb{R}^n$ . Moreover, we get the same for all series of partial derivatives, since we can always write them as

$$\sum_{k=0}^{r-1} a_k \frac{\partial^r f_k}{\partial x_{i_1} \cdots \partial x_{i_r}} + \sum_{k=r}^{\infty} a_k \frac{\partial^r f_k}{\partial x_{i_1} \cdots \partial x_{i_r}},$$

where the first part is a smooth function as it is a finite sum of smooth functions, and the second part can again be bounded from above by an absolutely converging series of numbers, so this expression will converge uniformly to  $\frac{\partial^r f}{\partial x_{i_1} \cdots \partial x_{i_r}}.$

It is apparent from the definition that the function  $f(x)$  satisfies the conditions of the lemma.  $\square$

#### PARTITION OF UNITY ON A MANIFOLD

**Theorem.** Consider a manifold  $M \subset \mathbb{R}^n$  equipped with a locally finite cover by open images  $U_i$  of parametrizations  $\varphi_i$ . Then, there exists a system of smooth non-negative functions  $f_i$  on the sets  $M$  such that for every point  $x \in M$ , we have  $\sum_i f_i(x) = 1$ , and  $f_i(x) \neq 0$  if and only if  $x \in U_i$ .

The system of functions  $f_i$  from the theorem is called the *partition of unity subordinated to the locally finite cover of the manifold* by the open sets  $U_i$ .

**PROOF.** First, we extend the sets  $U_i$  to open sets  $\tilde{U}_i$  using the extended parametrizations  $\tilde{\varphi}$ , from the definition of manifold and its local parametrizations. We can surely do this in such a way that the sets  $\tilde{U}_i$  keep being a locally finite cover of an open neighborhood  $\tilde{U} = \cup_i \tilde{U}_i \subset \mathbb{R}^n$  of the manifold  $M$ .

For every open set  $\tilde{U}_i$ , we can choose a non-negative function  $g_i(x)$  on the whole  $\mathbb{R}^n$  so that  $g_i(x) \neq 0$  exactly for  $x \in \tilde{U}_i$ . This can be done by Whitney's theorem proved in the above Lemma. Now, the function  $g(x) = \sum_i g_i(x)$  is well-defined for all  $x \in \mathbb{R}^n$  and smooth, thanks to the cover being locally finite (for every fixed point  $x$ , it is a finite sum of non-zero functions on some of its neighborhoods). The function  $g(x)$  is positive for all  $x \in M$ . Thus, instead of functions  $g_i(x)$  restricted to  $M$ , we may rather consider the functions  $f_i(x) = g_i(x)/g(x)$ , which already have both of the required properties of the theorem.  $\square$

**9.1.9. Integration of  $k$ -forms on manifolds.** Now, we are

ready for the definition of the integral of  $k$ -forms on  $k$ -dimensional manifolds. Let us consider an oriented manifold  $M \subset \mathbb{R}^n$  and a form  $\omega \in \Omega^k(M)$  with compact support.

Let us choose a locally finite cover of the manifold  $M$  by parametrizations  $\varphi_i : V_i \rightarrow U_i$  such that the closures of all images  $\varphi_i(V_i)$  are compact and, eventually, choose a partition of unity  $f_i$  subordinated to this cover.

The integral is defined by the formula

$$\int_M \omega = \int_M \sum_i f_i \omega = \sum_i \int_{U_i} f_i \omega,$$

and we get  $y^2 - x^2 = \sigma^2$ ,  $x + y = \sigma e^t$ ,

$$\sigma = \sqrt{y^2 - x^2}, \quad e^t = \frac{x + y}{\sqrt{y^2 - x^2}}.$$

$$u(0) = C_3 = f(\sigma) = \sqrt{1 + \sigma^2},$$

$$u(x, y) = u(\sigma(x, y), t(x, y)) = \sqrt{1 + \sigma^2} e^{2t}$$

Finally

$$u(x, y) = \sqrt{1 + y^2 - x^2} \cdot \frac{x + y}{y - x}, \text{ for } y > x.$$

Make a test.

How to write general solution of equation? Take an implicit description of characteristics, from  $x = C_1 e^t + C_2 e^t$ ,  $y = C_1 e^t - C_2 e^t$ ,  $u = C_3 e^{2t}$  we get

$$\frac{(x + y)^2}{u} = \frac{C_1^2}{C_3} = \Phi, \quad y^2 - x^2 = 2C_1 C_2 = \Psi.$$

General solution of equation is any function constant on characteristics:

$$u : \alpha(\Phi, \Psi) = 0, \text{ or } u = (x + y)^2 K(\Psi) = (x + y)^2 K(y^2 - x^2).$$

Make a test. In the next example we show another way, how to solve semilinear equation.  $\square$

**9.D.7.** Solve equation (S) and equation (L)

$$(S) \quad x^2 u_x + xy u_y = u^2, \quad (L) \quad x^2 u_x + xy u_y = x^2$$

both generally and for boundary conditions (i) — (iv)

- (i)  $u(1, y) = y$ ,
- (ii)  $u(\cos \sigma, \sin \sigma) = 1$ ,  $\sigma \in (-\frac{\pi}{2}, \frac{\pi}{2})$ ,
- (iii)  $u(x, 1 - x) = x$ ,
- (iv)  $u(x, 1 - x) = x^2$ .

**Solution.** Let's start again with characteristic equation

$$y' = \frac{dy}{dx} = \frac{xy}{x^2} \implies \frac{y}{x} = C.$$

Take a new coordinates  $\xi(x, y) = \frac{y}{x}$ ,  $\eta(x, y) = y$  and compute  $u_x = u_\xi \xi_x + u_\eta \eta_x = -\frac{y}{x^2} u_\xi$ ,  $u_y = u_\xi \xi_y + u_\eta \eta_y = -\frac{1}{x} u_\xi + u_\eta$ . The equation (S) will be now

$$u_\eta = \frac{u^2}{\eta^2} \cdot \xi$$

and solution

$$\int \frac{\partial u}{u^2} = \int \xi \frac{\partial \eta}{\eta^2} \implies -\frac{1}{u} = -\frac{\xi}{\eta} + K(\xi),$$

We can write general solution

$$u(\xi, \eta) = \frac{\eta}{\xi + \eta D(\xi)}, \quad u(x, y) = \frac{x}{1 + x D(\frac{y}{x})}.$$

Make a test. Now we can use general solution to find function  $D$ , depending on given boundary condition:

- (i)  $u(1, \sigma) = \frac{1}{1 + D(\frac{\sigma}{1})} = \sigma \implies D(\sigma) = \frac{1}{\sigma} - 1$ . From  $\frac{y}{x} = \frac{\sigma}{1}$  we get solution  $u(x, y) = \frac{x}{1 + x(\frac{x}{y} - 1)} = \frac{xy}{y + x^2 - xy}$ .

where the right-hand integrals have already been defined since each of the forms  $f_i \omega$  has support inside the image under the parametrization  $\varphi_i$  (and they equal to  $\int_M f_i \omega$  for the same reason).

Actually, we can assume that our sum is finite, since it suffices to consider integral over the image of parametrizations covering the compact support of  $\omega$ . Hence, it is a well-defined number, yet it remains to verify that the resulting value is independent of all our choices.

To this purpose, let us choose another system of parametrizations  $\psi : \tilde{V}_j \rightarrow \tilde{U}_j$ , again with compatible orientations, providing a locally finite cover of  $M$ . Let  $g_i$  be the corresponding partition of unity. Then the sets  $W_{ij} = U_i \cap \tilde{U}_j$  form again a locally finite covering and the set of functions  $f_i g_j$  provide the partition of unity subordinated to this covering. We arrive at the following equalities:

$$\begin{aligned} \sum_i \int_M f_i \omega &= \sum_i \int_M f_i (\sum_j g_j) \omega = \sum_{i,j} \int_M f_i g_j \omega \\ \sum_j \int_M g_j \omega &= \sum_j \int_M g_j (\sum_i f_i) \omega = \sum_{i,j} \int_M f_i g_j \omega, \end{aligned}$$

where the potentially infinite sums inside of the integrals are all locally finite, while the sums outside of the integral can be viewed as finite due to the compactness of the support of  $\omega$ . Thus, we have checked that the choices of the partition of unity and the parametrizations do not influence the value of the integral.

**9.1.10. Exterior differential.** As we have seen, the differential of a function can be interpreted as a mapping  $d : \Omega^0(\mathbb{R}^n) \rightarrow \Omega^1(\mathbb{R}^n)$ .



By means of parametrizations, this definition extends (in a coordinate free way) to functions  $f$  on manifolds  $M$ , where the differential  $df$  is a linear form on  $M$ . The following theorem extends this differential to arbitrary exterior forms on manifolds  $M \subset \mathbb{R}^n$ .

EXTERIOR DIFFERENTIAL

**Theorem.** For all  $m$ -dimensional manifolds  $M \subset \mathbb{R}^n$  and  $k = 0, \dots, m$ , there is the unique mapping

$$d : \Omega^k(M) \rightarrow \Omega^{k+1} M,$$

such that

- (1)  $d$  is linear with respect to multiplication by real numbers;
- (2) for  $k = 0$ , this is the differential of functions;
- (3) if  $\alpha \in \Omega^k(M)$ ,  $\beta$  arbitrary, then  $d(\alpha \wedge \beta) = (d\alpha) \wedge \beta + (-1)^k \alpha \wedge (d\beta)$ ;
- (4)  $d(df) = 0$  for every function  $f$  on  $M$ .

The mapping  $d$  is called the exterior differential. The equality  $d \circ d = 0$  is valid for all degrees  $k$ .

(ii)  $u(\cos \sigma, \sin \sigma) = \frac{\cos \sigma}{1 + \cos \sigma} = 1$  we have  $D(\sigma) = 1 - \frac{1}{\cos \sigma}$  and because  $\sigma = \arctan \frac{y}{x}$ ,  $\cos \sigma = \frac{1}{\sqrt{1 + \frac{y^2}{x^2}}}$  so

$D(x, y) = 1 - \sqrt{1 + \frac{y^2}{x^2}}$  and the solution is

$$u(x, y) = \frac{x}{1 + x \left[ 1 - \sqrt{1 + \frac{y^2}{x^2}} \right]}, \quad x > 0.$$

(iii) Analogically  $u(x, y) = x$ ,

(iv) Analogically  $u(x, y) = \frac{x^2}{x + y^2 + xy}$ .

For linear equation (L) is situation much easier. We can see, that  $u = x$  is one of the solutions of both, (L) and (S). For (L) we can use principle of superposition and write general solution like a sum of general solution of homogenous case and any particular solution (make a test):

$$u(x, y) = x + D\left(\frac{y}{x}\right).$$

For boundary conditions we get

(i)  $u(x, y) = x + \frac{y}{x} - 1$ ,

(ii)  $u(x, y) = x + 1 - \frac{x}{\sqrt{x^2 + y^2}}$ ,  $x > 0$ ,

(iii)  $u(x, y) = x$ ,

(iv)  $u(\sigma, 1 - \sigma) = \sigma + D(\sigma) = \sigma^2 \Rightarrow D(\sigma) = \sigma^2 - \sigma$ ,  $\frac{1 - \sigma}{\sigma} = \frac{y}{x} \Rightarrow \sigma = \frac{x}{x + y}$ ,  $D = \sigma^2 - \sigma = -\frac{xy}{(x + y)^2}$  and solution is

$$u(x, y) = x - \frac{xy}{(x + y)^2}, \quad x \neq -y.$$

Make a test for all solutions and use this method also for solving previous problem.  $\square$

**9.D.8.** Find general solution of

$$yu_x + u_y = -u.$$

**9.D.9.** Find general solution of

$$u_x + u_y = -u(x - y)$$

and solutions for given boundary condition

(i)  $u(0, y) = y$ ,

(ii)  $u(\sigma, -\sigma) = \frac{1}{\sigma}$ ,

(iii)  $u(x, 0) = x^2$ .

**9.D.10.** Solve

$$u_x - y^2 u_y = u^2, \quad u(x, 1) = \frac{1}{2x}.$$

**9.D.11.** Solve

$$xu_x + yu_y = 2(x^2 + y^2), \quad u(1, y) = 2y^2 + 1.$$

**9.D.12.** Solve

$$2xu_x - yu_y = x^2 + y^2, \quad u(2, y) = 1 - y^2.$$

**PROOF.** Each  $k$ -form can be written locally in the form

$$\alpha = \sum_{i_1 < \dots < i_k} a_{i_1 \dots i_k} dx_{i_1} \wedge \dots \wedge dx_{i_k}.$$

If the differential  $d$  exists, then by the required properties, it must be equal to

$$\begin{aligned} d\alpha &= \sum_{i_1 < \dots < i_k} da_{i_1 \dots i_k} \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k} \\ (5) \quad &= \sum_{j; i_1 < \dots < i_k} \frac{\partial a_{i_1 \dots i_k}}{\partial x_j} dx_j \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k}. \end{aligned}$$

Indeed, the generators  $dx_i$  of linear forms are in fact the differentials of the coordinate functions, so further differentiation must lead to zero by the last property, while we know the differential of functions. Further, we have  $d(f\beta) = df \wedge \beta + f d\beta$  by property (3).

Thus, let us define the differential  $d$  in coordinates by the formula (5), and we are going to verify all of the required properties. We shall proceed in two steps.

First, we check the requirements in one coordinate patch. The first two requirements are obvious from the formula (5). It is enough to verify the property (3) for the special forms  $\alpha = a dx_{i_1} \wedge \dots \wedge dx_{i_k}$  and  $\beta = b dx_{j_1} \wedge \dots \wedge dx_{j_\ell}$ , since then the property must hold for the sums of such forms too.

We compute

$$\begin{aligned} d(\alpha \wedge \beta) &= d(ab dx_{i_1} \wedge \dots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \dots \wedge dx_{j_\ell}) \\ &= \sum_i \frac{\partial a}{\partial x_i} b dx_i \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \dots \wedge dx_{j_\ell} \\ &\quad + \sum_i \frac{\partial b}{\partial x_i} a dx_{i_1} \wedge \dots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \dots \wedge dx_{j_\ell} \\ &= d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta, \end{aligned}$$

as expected.

The last property can be again verified for simple forms  $\alpha = a dx_{i_1} \wedge \dots \wedge dx_{i_k}$ . Applying the formula (5) twice, we arrive at

$$\begin{aligned} d(d\alpha) &= d\left(\sum_i \frac{\partial a}{\partial x_i} dx_i \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k}\right) \\ &= \sum_j \sum_i \frac{\partial^2 a}{\partial x_i \partial x_j} dx_j \wedge dx_i \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k} \\ &= \sum_{i < j} \left(\frac{\partial^2 a}{\partial x_i \partial x_j} - \frac{\partial^2 a}{\partial x_j \partial x_i}\right) dx_j \wedge dx_i \wedge dx_{i_1} \wedge \dots \wedge dx_{i_k} \\ &= 0, \end{aligned}$$

where we used the fact that the wedge product  $dx_i \wedge dx_i$  vanishes,  $dx_j \wedge dx_i = -dx_i \wedge dx_j$  and the second partial derivatives of functions are symmetric.

9.D.13. Solve

$$x^2 u_x + y u_y = 2u, \quad u(1, y) = y^3.$$

9.D.14. Solve equation for given boundary condition

$$u = u_x u_y, \quad u(x, 0) = x^2.$$

**Solution.** We show, how to use method of characteristics for general first order equation

$$F(p, q, u, x, y) = 0, \quad \text{where } p = u_x, q = u_y.$$

Our equation we can write in the form  $u - p \cdot q = 0$ . We need to solve system of five ODE's:

$$\begin{aligned} \dot{x} &= F_p = -q, \\ \dot{y} &= F_q = -p, \\ \dot{u} &= pF_p + qF_q = -2pq, \\ \dot{p} &= -F_x - pF_u = -p, \\ \dot{q} &= -F_y - qF_u = -q. \end{aligned}$$

Solution of this characteristic system is

$$\begin{aligned} x &= C_1 e^{-t} + C_2, \\ y &= C_3 e^{-t} + C_4, \\ u &= C_1 C_2 e^{-2t} + C_5, \\ p &= C_3 e^{-t}, \\ q &= C_1 e^{-t}. \end{aligned}$$

Without lose of generality take  $t = 0$  for the point of intersection of characteristic lines with boundary line  $x = \varphi(\sigma) = \sigma, y = \psi(\sigma) = 0, u = f(\sigma) = \sigma^2$  and find  $C_i = C_i(\sigma)$  from

$$\begin{aligned} x(0) &= C_1 + C_2 = \sigma, \\ y(0) &= C_3 + C_4 = 0, \\ u(0) &= C_1 C_2 + C_5 = \sigma^2, \\ p(0) &= C_3 = p_0 = 2\sigma, \\ q(0) &= C_1 = q_0 = \frac{\sigma}{2}. \end{aligned}$$

The value of  $p_0$  and  $q_0$  we get from system of two algebraic equation:

$$F(p_0, q_0, f(\sigma), \varphi(\sigma), \psi(\sigma)) = 0 \implies \sigma^2 = p_0 \cdot q_0,$$

$$u(\varphi(\sigma), \psi(\sigma)) = f(\sigma) \implies \frac{d}{d\sigma} u(\varphi(\sigma), \psi(\sigma)) = \frac{d}{d\sigma} f(\sigma),$$

$$u_x \varphi'(\sigma) + u_y \psi'(\sigma) = f'(\sigma) \implies p_0 \cdot 1 + q_0 \cdot 0 = 2\sigma.$$

Finally we get  $C_1 = \frac{\sigma}{2}, C_2 = \frac{\sigma}{2}, C_3 = 2\sigma, C_4 = -2\sigma, C_5 = 0$ . From

$$x(\sigma, t) = \frac{\sigma}{2} (e^{-t} + 1), \quad y(\sigma, t) = 2\sigma (e^{-t} - 1)$$

we can write

$$\sigma(x, y) = x - \frac{y}{4}, \quad e^{-t}(x, y) = \frac{4x + y}{4x - y},$$

○



○

The second step in the proof is the verification, that the coordinate formula (5) correctly defines a differential operator on general manifolds  $M$ . In order to achieve this, it is sufficient to show that the coordinate expression of the exterior derivative commutes with the pullbacks of forms. Indeed, we may then define the differential operator around any point in any coordinates and the results will coincide.<sup>3</sup>

Thus, consider a change of coordinates  $G : U \rightarrow V, x = G(y) = (g_1(y), \dots, g_m(y))$ , and compute  $G^*(d\alpha)$  of an exterior form  $\alpha = a dx_{i_1} \wedge \dots \wedge dx_{i_k}$  (which gives the result for sums of such expressions too). This is straightforward:

$$\begin{aligned} G^*(d\alpha) &= \sum_i G^*\left(\frac{\partial a}{\partial x_i}\right) G^*(dx_{i_1}) \wedge G^*(dx_{i_2}) \wedge \dots \\ &= \sum_i \left(\frac{\partial a}{\partial x_i} \circ G\right) \left(\frac{\partial g_i}{\partial y_1} dy_1 + \dots\right) \wedge \left(\frac{\partial g_{i_1}}{\partial y_1} dy_1 + \dots\right) \wedge \dots \end{aligned}$$

Now, notice  $d(G^*(dx_j)) = G^*(d(dx_j)) = 0$  and thus

$$\begin{aligned} d(G^*(\alpha)) &= d((a \circ G) G^*(dx_{i_1}) \wedge \dots) \\ &= d(a \circ G) \wedge G^*(dx_{i_1}) \wedge \dots \wedge G^*(dx_{i_k}) \\ &= \sum_i \left(\left(\frac{\partial a}{\partial x_i} \circ G\right) \left(\frac{\partial g_i}{\partial y_1} dy_1 + \dots\right)\right) \wedge G^* dx_{i_1} \wedge \dots, \end{aligned}$$

clearly the same expressions.  $\square$

9.1.11. **Manifolds with boundary.** In practical problems,



we often work with manifolds  $M$  like an open ball in the three-dimensional space. At the same time, we are interested in the boundaries of these manifolds  $\partial M$ , which is a sphere in the case of a ball.

The simplest case is the one of connected curves. It is either a closed curve (like a circle in the plane), then its boundary is empty, or the boundary is formed by two points. These points will be considered including the orientation inherited from the curve, i.e. the initial point will be taken with the minus sign, and the terminal point with the plus sign.

The curve integral is the easiest one, and we can notice that integrating the differential  $df$  of a function along the curve  $M$  defined as the image of a parametrization  $c : [a, b] \rightarrow M$ , then we get directly from the definition that

$$\begin{aligned} \int_M df &= \int_{[a,b]} c^*(df) = \int_a^b \frac{d}{dt} (f \circ c)(t) dt \\ &= f(c(b)) - f(c(a)). \end{aligned}$$

Therefore, the result is not only independent of the selected parametrization, but also of the actual curve. Only the initial and terminal points matter. Splitting the curve into several

<sup>3</sup>Such operators are intrinsically defined on all manifolds. Actually, for all  $k > 0$ , the only operation  $d : \Omega^k \rightarrow \Omega^{k+1}$  commuting with pullbacks and with values depending only on the behavior of the argument  $\alpha$  on any small neighborhood of  $x$  (locality of the operator), is the exterior derivative. Thus even the linearity, as well as the dependence on the first derivatives are direct consequences of naturality. See the book *Natural operations in differential geometry*, Springer, 1993, by I. Kolar, P.W. Michor and J. Slovák for full proof of this astonishing claim.

and

$$u(\sigma(x, y), t(x, y)) = \sigma^2 e^{-2t} = \left[ \left( x - \frac{y}{4} \right) \frac{4x + y}{4x - y} \right]^2,$$

$$u(x, y) = \frac{1}{16} (4x + y)^2.$$

Make a test.

**9.D.15.** Solve

$$u_x^2 + u_y^2 = 1, \quad u(\cos \sigma, \sin \sigma) = 1.$$

**9.D.16.** Solve

$$u_x^2 + u_y^2 + 2u = 0, \quad u(\cos \sigma, \sin \sigma) = -\frac{1}{2}.$$

**9.D.17.** Solve

$$\sqrt{u_x^2 + u_y^2} - u = 0, \quad u(\cos(\sigma), \sin(\sigma)) = 1.$$

**9.D.18.** Solve

$$u_x^2 + u_y^2 - u = 0, \quad u(\cos(\sigma), \sin(\sigma)) = 1.$$

**9.D.19.** Solve

$$\sqrt{u_x^2 + u_y^2} + u = 0, \quad u(\cos(\sigma), \sin(\sigma)) = -1.$$

**9.D.20.** Solve

$$u_x u_y + u = 0, \quad u(x, 0) = x.$$

**9.D.21.** Solve

$$x u_x^2 + y u_y^2 = u, \quad u(\sigma, \sigma) = 2\sigma.$$

**9.D.22.** Solve

$$x u_x^2 + y u_y^2 = u, \quad u(1, y) = 1.$$

**9.D.23.** Solve

$$u_x^2 - u_y^2 = 4u, \quad u(\cos \sigma, \sin \sigma) = \cos 2\sigma.$$

**9.D.24.** Solve

$$u_x^2 + y u_y = u, \quad u(1, y) = y.$$

consecutive disjoint intervals, the integral splits into the sum of differences of the values at the splitting points. This sum will be telescoping (i.e., the middle terms cancel out), resulting in the same value again.

Notice, we have already proved the behavior expected in 9.1.5 when dealing with the elevation gain by a cyclist.

- We shall discuss this phenomenon in general dimensions now. To be able to do this, we need to formalize the concept of the boundary of a manifold and its orientation. The simplest case is the closed half-space  $\bar{M} = (-\infty, 0] \times \mathbb{R}^{n-1}$ . Its boundary is  $\partial M = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n; x_1 = 0\}$ .
- The orientation on this boundary inherited from the standard orientation is the one determined by the form  $dx_2 \wedge \dots \wedge dx_n$ .

ORIENTED BOUNDARY OF A MANIFOLD

- Let us consider a closed subset  $\bar{M} \subset \mathbb{R}^n$  such that its interior  $M \subset \bar{M}$  is an oriented  $m$ -dimensional manifold covered by compatible parametrizations  $\varphi_i$ . Further, let us assume that for every boundary point  $x \in \partial M = \bar{M} \setminus M$ , there is a neighborhood in  $\bar{M}$  with parametrization  $\varphi : V \subset (-\infty, 0] \times \mathbb{R}^{m-1} \rightarrow M$  such that the points  $x \in \partial M \cap \varphi(V)$  from just the image of the boundary of the half-space  $(-\infty, 0] \times \mathbb{R}^{m-1}$ . The subset  $\bar{M} \subset \mathbb{R}^m$  covered by the above parametrizations with compatible orientations is called an *oriented manifold with boundary*.
- The restrictions of the parametrizations including boundary points to the boundary  $\partial M$  defines the structure of an  $(m - 1)$ -dimensional oriented manifold on  $\partial M$ .

- Think of a closed unit balls  $B(x, r) \subset \mathbb{R}^n$  as such manifolds. Their interiors are an  $n$ -dimensional manifolds, just open subsets in  $\mathbb{R}^n$ , but their boundaries  $S^{n-1}$  are the spheres with the inherited structure of  $(n-1)$ -dimensional manifolds. The inherited orientations are well understood via the outward normals to the spheres. Another example is a plane disc sitting as a 2-dimensional manifold in  $\mathbb{R}^3$  with its 1-dimensional boundary being a circle. Here the chosen position of the normal to the plane defines the orientation of the circle, one or the other way.

pridat obrazek k prikladu

- In practice, we often deal with slightly more general manifolds where we allow for corners in the boundary of all smaller dimensions. A good example is the cube in  $\mathbb{R}^3$  having the sides as 2-dimensional parts of the boundary and also the edges between them as 1-dimensional parts and the vortices as 0-dimensional parts of the boundary. Yet another class of examples is formed by all simplexes and their curved embeddings in  $\mathbb{R}^n$ . Since those lower dimensional parts of the boundary will have Riemann measure zero, we can neglect them when integrating over  $\partial M$ . Thus we shall not go into details of this technical extension of our definitions.

○

**9.D.25.** Find all solutions (if exists) of the system

$$\begin{aligned} u_x &= f(x, y, u) = y(u - xy + 1), \\ u_y &= g(x, y, u) = x(u - xy + 1). \end{aligned}$$

**Solution.** Necessary (and sufficient) condition for existence of solution is  $u_{xy} = u_{yx}$ :

$$\begin{aligned} u_{xy} &= f_y + f_u g = u - 2xy + 1 + xyu - x^2y^2 + xy, \\ u_{yx} &= g_x + g_u f = u - 2xy + 1 + xyu - x^2y^2 + xy. \end{aligned}$$

Let's start with first equation

$$\frac{\partial u}{\partial x} = yu - xy^2 + 1,$$

which is linear and we can find solution in the form

$$u(x, y) = K(x, y)e^{xy}$$

so

$$u_x = K_x e^{xy} + yK e^{xy} = yK e^{xy} - xy^2 + y$$

and

$$K_x = (y - xy^2)e^{-xy},$$

by per partes we have

$$K(x, y) = xy e^{-xy} + D(y).$$

Substituing this to the second equation

$$u(x, y) = xy + D(y)e^{xy} \implies$$

$$u_y = x + D'(y)e^{xy} + xD(y)e^{xy} = g(x, y, u)$$

we get  $D'(y) = 0$ ,  $D$  is constant and all solutions are given by

$$u(x, y) = xy + De^{xy}.$$

□

**9.D.26.** Find solution (if exists) of system

$$u_x = 2xy^2u, \quad u_y = 2x^2yu.$$

○

**9.D.27.** Find solution (if exists) of system

$$xu_x = u - y, \quad yu_y = u - x.$$

○

**9.D.28.** For second order semilinear equation we will use notation

$$A(x, y)u_{xx} + 2B(x, y)u_{xy} + C(x, y)u_{yy} = F(u_x, u_y, u, x, y).$$

Show, that characteristics polynomial of the matrix

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix}$$

has

- (E) two nonzero real roots (not necessarily different) of the same sign ( $\lambda_1 \cdot \lambda_2 > 0$ ) if and only if  $B^2 - AC < 0$ , equation is elliptic,
- (H) two nonzero real roots of different sign ( $\lambda_1 \cdot \lambda_2 < 0$ ) if and only if  $B^2 - AC > 0$ , equation is hyperbolic,
- (P) two real roots from which (at least) one is zero ( $\lambda_1 \cdot \lambda_2 = 0$ ) if and only if  $B^2 - AC = 0$ , equation is parabolic.

**9.1.12. Stokes' theorem.** Now, we get to a very important and useful result. We shall formulate the main theorem about the multidimensional analogy of curve integrals for smooth forms and smooth manifolds. A brief analysis of the proof shows that actually, we need once continuously differentiable exterior forms as integrands on twice continuously differentiable parametrizations of the manifold.



In practice, the boundary of the region is often similar as in the case of the unit cube in  $\mathbb{R}^3$ , i.e., we have discontinuities of the derivatives on a Riemann measurable set with measure zero in the boundary. In such a case, we divide the integration to smooth parts and add the results up. We can notice that although new pieces of boundaries appear, they are adjacent and have opposite orientations in the adjacent regions, so their contribution is canceled out (just like in the above case of boundary points of a piecewise differentiable curve).

#### STOKES' THEOREM

**Theorem.** Consider a smooth exterior  $(k-1)$ -form  $\omega$  with compact support on an oriented manifold  $\bar{M}$  with boundary  $\partial M$  with the inherited orientation. Then we have

$$\int_M d\omega = \int_{\partial M} \omega.$$

**PROOF.** Using an appropriate locally finite cover of the manifold  $\bar{M}$  and a partition of unity subordinated to it, we can express the integrals on both sides as the sum (even a finite sum, since the support of the considered form  $\omega$  is compact) of integrals of forms  $\omega$  supported in individual parametrizations. Thus we can restrict ourselves to just two cases  $\bar{M} = \mathbb{R}^k$  or the half-space  $\bar{M} = (-\infty, 0] \times \mathbb{R}^{k-1}$ .



In both cases,  $\omega$  will surely be the sum of forms  $\omega_j$

$$\omega_j = a_j(x) dx_1 \wedge \cdots \wedge \hat{dx}_j \wedge \cdots \wedge dx_k,$$

where the hat indicates the omission of the corresponding linear form, and  $a_j(x)$  is a smooth function with compact support. Their exterior differentials are

$$d\omega_j = (-1)^{j-1} \frac{\partial a_j}{\partial x_j} dx_1 \wedge \cdots \wedge dx_k.$$

Again, we can verify the claim of the theorem for such forms  $\omega_j$  separately. Let us compute the integrals  $\int_M d\omega_j$  using the Fubini's theorem. This is most simple if  $\bar{M} = \mathbb{R}^n$ ,

$$\begin{aligned} \int_{\mathbb{R}^n} d\omega_j &= (-1)^{j-1} \int_{\mathbb{R}^{k-1}} \left( \int_{-\infty}^{\infty} \frac{\partial a_j}{\partial x_j} dx_j \right) dx_1 \cdots \hat{dx}_j \cdots dx_k \\ &= (-1)^{j-1} \int_{\mathbb{R}^{k-1}} [a_j]_{-\infty}^{\infty} dx_1 \cdots \hat{dx}_j \cdots dx_k = 0. \end{aligned}$$

Notice, we are allowed to use the Fubini's theorem for the entire  $\mathbb{R}^n$  since the support of the integrated function is in fact compact and thus we can replace the integration domain by a large multidimensional interval  $I$ . At the same time, the forms  $\omega_j$  are all zero outside of such a large interval  $I$  and thus

**Solution.** EASY ale FUJ

**9.D.29.** Show, that solution of

(H) hyperbolic equation in canonical form

$$u_{xy} = 0$$

is any function  $u(x, y) = F(x) + G(y)$ .

(E) elliptic equation in canonical form

$$\Delta v = v_{xx} + v_{yy} = 0$$

is any function  $v(x, y) = F(x + iy) + G(x - iy)$ .

(P) parabolic equation in canonical form

$$w_{xx} = kw_y,$$

is any function

$$w(x, y) = e^{k(Cx+C^2y)}.$$

**Solution.**

$$u_x = F', u_{xy} = 0.$$

$$v_x = F' + G', v_{xx} = F'' + G'', v_y = iF' - iG',$$

$$v_{yy} = -F'' - G'' \implies u_{xx} + u_{yy} = 0,$$

$$w_x = kCe^{k(Cx+C^2y)}, w_{xx} = k^2C^2e^{k(Cx+C^2y)},$$

$$w_y = kC^2e^{k(Cx+C^2y)} \implies w_{xx} = kw_y.$$

**9.D.30.** Show, that real solution of Laplace equation in polar coordinates  $x = r \cos \varphi, y = r \sin \varphi$  (compute):

$$u_{rr} + \frac{1}{r^2}u_{\varphi\varphi} + \frac{1}{r}u_\varphi = 0,$$

are *harmonic* function given by

$$\alpha_n(r, \varphi) = r^n \cos n\varphi, \beta_n(r, \varphi) = r^n \sin n\varphi, n \in \mathbf{N}_0.$$

Find their expression  $\alpha_n(x, y), \beta_n(x, y)$  for  $n = 0, 1, 2, 3$ .

○

**9.D.31.** Find general solution of second order equation

$$x^2u_{xx} - 2xyu_{xy} + y^2u_{yy} - x^2u_x + (x+2)yu_y = 0.$$

**Solution.** We have

$$A = x^2, B = -xy, C = y^2, B^2 - AC = 0$$

and equation is parabolic. Let's solve characteristic equation

$$\frac{dy}{dx} = y' = \frac{B \pm \sqrt{B^2 - AC}}{A} = \frac{-xy}{x^2} = -\frac{y}{x}.$$

We have solution in implicit form  $xy = C = \xi$  ( $\xi = \xi(x, y)$  will be new coordinate function). For second new coordinate function  $\eta = \eta(x, y)$  we can take any independent ( $\xi_x \eta_y \neq \xi_y \eta_x$ ) function, for example  $\eta = x$ . Inverse transformation is

$$x = \eta, y = \frac{\xi}{\eta}.$$

We get

$$u_x = u_\xi \xi_x + u_\eta \eta_x = \frac{\xi}{\eta} u_\xi + u_\eta, u_y = u_\xi \xi_y + u_\eta \eta_y = \eta u_\xi,$$

□ the integrals  $\int_{\partial M} \omega_j$  all vanish and the claim of the Stokes' theorem is verified in this case. Actually, we may also say that  $\partial M = \emptyset$  and thus the integral is zero.

Next, let us assume  $\bar{M}$  is the half-space  $(-\infty, 0] \times \mathbb{R}^{k-1}$ . If  $j > 1$ , the form  $\omega_j$  evaluates identically to zero on the boundary  $\partial M$ , since  $x_1$  is constant there and thus  $dx_1$  is identically zero on all tangent directions to  $\partial M$ . Integration over the interior  $M$  yields zero, using the same approach as above:

$$\begin{aligned} \int_M d\omega_j &= (-1)^{j-1} \int_{-\infty}^0 \int_{\mathbb{R}^{k-2}} \left( \int_{-\infty}^{\infty} \frac{\partial a_j}{\partial x_j} dx_j \right) dx_1 \cdots \hat{dx}_j \cdots dx_k \\ &= (-1)^{j-1} \int_{-\infty}^0 \int_{\mathbb{R}^{k-1}} [a_j]_{-\infty}^{\infty} dx_1 \cdots \hat{dx}_j \cdots dx_k = 0 \end{aligned}$$

since the function  $a_j$  has compact support. So the theorem is also true in this case.

However, if  $j = 1$ , then we obtain

$$\begin{aligned} \int_M d\omega_1 &= \int_{\mathbb{R}^{k-1}} \left( \int_{-\infty}^0 \frac{\partial a_1}{\partial x_1} dx_1 \right) dx_2 \cdots dx_k \\ &= \int_{\mathbb{R}^{k-1}} a_1(0, x_2, \dots, x_k) dx_2 \cdots dx_k = \int_{\partial M} \omega_1. \end{aligned}$$

This finishes the proof of Stokes' theorem. □

**9.1.13. Green's theorem.** We have proved an extraordinarily strong result which covers several standard integral relations from the classical vector analysis. For instance, we can notice that by Stokes theorem, the integration of exterior differential  $d\omega$  of any  $k$ -form over a compact manifold without boundary is always zero (for example, integral of any 2-form  $d\omega$  over the sphere  $S^2 \subset \mathbb{R}^3$  vanishes).

Let us look step by step at the cases of Stokes' theorem with  $k$  dimensional boundaries  $\partial M$  in  $\mathbb{R}^n$  in low dimensions.

#### GREEN'S THEOREM

In the case  $n = 2, k = 1$ , we are examining a domain  $M$  in the plane, bounded by a closed curve  $C = \partial M$ . Differential 1-forms are  $\omega(x, y) = f(x, y) dx + g(x, y) dy$ , with the differential  $d\omega = \left(-\frac{\partial f}{\partial y} + \frac{\partial g}{\partial x}\right) dx \wedge dy$ . Therefore, Stokes' theorem yields the formula

$$\int_C f(x, y) dx + g(x, y) dy = \int_M \left(-\frac{\partial f}{\partial y} + \frac{\partial g}{\partial x}\right) dx \wedge dy$$

which is one of the standard forms of the *Green's theorem*.

Using the standard scalar product on  $\mathbb{R}^2$ , we can identify the vector field  $X$  with a linear form  $\omega_X$  such that  $\omega_X(Y) = \langle Y, X \rangle$ . In the standard coordinates  $(x, y)$ , this just means that the field  $X = f(x, y) \frac{\partial}{\partial x} + g(x, y) \frac{\partial}{\partial y}$  corresponds to the form  $\omega = f(x, y) dx + g(x, y) dy$  given above.

The integral of  $\omega_X$  over a curve  $C$  has the physical interpretation of the work done by movement along this curve in the force field  $X$ .

Green's theorem then says, besides others, that if  $\omega_X = dF$  for some function  $F$ , then the work done along a closed curve is always zero. Such fields are called potential fields



$$u_{xx} = \frac{\xi^2}{\eta^2} u_{\xi\xi} + 2\frac{\xi}{\eta} u_{\xi\eta} + u_{\eta\eta},$$

$$u_{xy} = \xi u_{\xi\xi} + \eta u_{\xi\eta} + u_{\xi}, u_{yy} = \eta^2 u_{\xi\xi}.$$

Finally the equation has the form

$$\eta^2(u_{\eta\eta} - u_{\eta}) = 0 \implies u_{\eta\eta} = u_{\eta}.$$

Integrating twice we have  $u(\xi, \eta) = G(\xi)e^{\eta} + F(\xi)$ ,

$$u(x, y) = G(xy)e^x + F(xy), \text{ make a test.}$$

Try to solve equation for different choice of the function  $\eta$ .

□

**9.D.32.** Find general solution of second order equation

$$y u_{yy} - x u_{xy} + u_y = 0.$$

**Solution.** We have

$$A = 0, B = -\frac{x}{2}, C = y, B^2 - AC = \frac{x^2}{4} > 0 \text{ for } x \neq 0$$

and this equation is hyperbolic. The characteristic equation (we can't divide by  $A = 0!$ ) is now

$$\frac{dx}{dy} = x' = \frac{B \pm \sqrt{B^2 - AC}}{C} = \frac{1}{y} \left( -\frac{x}{2} \pm \frac{x}{2} \right).$$

We have two solutions in implicit form (new coordinate functions  $\xi = \xi(x, y), \eta = \eta(x, y)$ )

$$x = C_1 = \xi, \quad xy = C_2 = \eta \implies x = \xi, y = \frac{\eta}{\xi}.$$

Finally

$$u_x = u_{\xi} + \frac{\eta}{\xi} u_{\eta}, \quad u_y = \xi u_{\eta},$$

$$u_{xy} = \xi u_{\xi\eta} + \eta u_{\eta\eta} + u_{\eta}, \quad u_{yy} = \xi^2 u_{\eta\eta}$$

and in new coordinates  $\xi, \eta$  has the equation canonical form

$$u_{\xi\eta} = 0.$$

General solution is

$$u(\xi, \eta) = F(\xi) + G(\eta), \quad u(x, y) = F(x) + G(xy).$$

□

**9.D.33.** Find general solution of second order equation

$$u_{xx} - 2 \sin x u_{xy} + (2 - \cos^2 x) u_{yy} - \cos x u_y = 0.$$

**Solution.** We have

$$A = 1, B = -\sin x, C = 2 - \cos^2 x, B^2 - AC < 0$$

and equation is elliptic. Let's solve characteristic equation

$$\frac{dy}{dx} = y' = \frac{B \pm \sqrt{B^2 - AC}}{A} = -\sin x \pm i.$$

We have solution in implicit form

$$y - \cos x - ix = C_1 = \Phi(x, y), \quad y - \cos x + ix = C_2 = \Psi(x, y)$$

and new (real) coordinate function are

$$\xi = \frac{1}{2}(\Phi + \Psi) = y - \cos x, \quad \eta = \frac{1}{2i}(\Psi - \Phi) = x.$$

and the function  $F$  is the potential of the field  $X$ . In other words, the work done when moving in potential fields does not depend on the path, it depends only on the initial and terminal points.

With Green's theorem, we have verified once again that integrating the differential of a function along a curve depends solely on the initial and terminal points of the curve.

**9.1.14. The divergence theorem.** The next case deals with integrating over some open subset in  $\mathbb{R}^3$  and it has got a lot of incarnations in practical use. We shall mention a few.

#### GAUSS–OSTROGRADSKY'S THEOREM

In the case  $n = 3, k = 2$  we are examining a region  $M \subset \mathbb{R}^3$ , bounded by a surface  $S$ . All 2-forms are of the form  $\omega = f(x, y, z) dy \wedge dz + g(x, y, z) dz \wedge dx + h(x, y, z) dx \wedge dy$ , and we get  $d\omega = \left( \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z} \right) dx \wedge dy \wedge dz$ .

The Stokes' theorem says that

$$\begin{aligned} \int_S f dy \wedge dz + g dz \wedge dx + h dx \wedge dy \\ = \int_M \left( \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z} \right) dx \wedge dy \wedge dz. \end{aligned}$$

This is the statement of the *Gauss–Ostrogradsky theorem*.

This theorem has a very illustrative physical interpretation, too.

Every vector field  $X = f(x, y, z) \frac{\partial}{\partial x} + g(x, y, z) \frac{\partial}{\partial y} + h(x, y, z) \frac{\partial}{\partial z}$  can be plugged into the first argument of the standard volume form  $\omega_{\mathbb{R}^3} = dx \wedge dy \wedge dz$  on  $\mathbb{R}^3$ . Clearly, the result is a 2-form  $\omega^X(x, y, z) = f(x, y, z) dy \wedge dz + g(x, y, z) dz \wedge dx + h(x, y, z) dx \wedge dy$ .

The latter 2-form infinitesimally describes the volume of the parallelepiped given by the flux caused by the field  $X$  through a linearized piece of surface. If we consider the vector field to be the velocity of the flow of the particular points of the space, this infinitesimally describes the volume transported pointwise by the flow through the given surface  $S$ . Thus the left hand side is the total change of volume inside of  $S$ , caused by the flow of  $X$ .

The integrand of the right-hand side of the integral, is related to the so-called *divergence of the vector field*, which is the expression defined as

$$d(\omega^X) = (\operatorname{div} X) dx \wedge dy \wedge dz.$$

The Gauss–Ostrogradsky theorem says

$$\int_S i_X \omega_{\mathbb{R}^3} = \int_M \operatorname{div} X \omega_{\mathbb{R}^3},$$

i.e. the volume of total flow through a surface is given as the integral of the divergence of the vector field over the interior. In particular, if  $\operatorname{div} X$  vanished identically, then the total volume flow through the boundary surface of the region is zero as well.

We have

$$\xi_x = \sin x, \xi_y = 1, \xi_{xx} = \cos x, \xi_{xy} = 0, \xi_{yy} = 0,$$

$$\eta_x = 1, \eta_y = \eta_{xx} = \eta_{xy} = \eta_{yy} = 0,$$

$$u_y = u_\xi, u_{xy} = u_{\xi\xi} \sin \eta + u_{\xi\eta}, u_{yy} = u_{\xi\xi},$$

$$u_{xx} = u_{\xi\xi} \sin^2 \eta + 2u_{\xi\eta} \sin \eta + u_{\eta\eta} + u_\xi \cos \eta.$$

Finally the equation became canonical

$$\Delta u = u_{\xi\xi} + u_{\eta\eta} = 0.$$

General solution is

$$u(\xi, \eta) = C(\xi + i\eta) + D(\xi - i\eta) \text{ make a test,}$$

$$u(x, y) = C(y - \cos x + ix) + D(y - \cos x - ix).$$

Real solution is given by harmonic functions in variables  $\xi, \eta$ . □

**9.D.34.** Solve

$$x^2 u_{xx} - 2x u_{xy} + u_{yy} + x u_x = 0.$$

**9.D.35.** Solve

$$x u_{xx} - y u_{xy} + u_x = 0$$

**9.D.36.** Solve

$$u_{xx} - u_{xy} + 2u_{yy} = 0.$$

**9.D.37.** Find canonical form of

$$y^2 u_{xx} + x^2 u_{yy} = 0.$$

**9.D.38.** Show, that solution of nonhomogenous wave equation with initial conditions (infinite string):

$$u_{tt}(t, x) = a^2 u_{xx}(t, x) + f(t, x), \quad t \in (0, \infty), x \in \mathbf{R},$$

$$u(0, x) = \varphi(x), \quad u_t(0, x) = \psi(x), \quad x \in \mathbf{R},$$

is given by d'Alembert's formula:

$$u(t, x) = \frac{\varphi(x - at) + \varphi(x + at)}{2} + \frac{1}{2a} \int_{x-at}^{x+at} \psi(\xi) d\xi +$$

$$+ \frac{1}{2a} \int_0^t \left( \int_{x-a(t-\sigma)}^{x+a(t-\sigma)} f(\sigma, \xi) d\xi \right) d\sigma.$$

Such fields, with  $\operatorname{div} X = 0$ , are called *divergence free* or *solenoidal* vector fields. They correspond to dynamics without changes of volumes (e.g. modelling dynamics of incompressible liquids).

In order to reformulate the theorem completely in terms of functions, let us observe that the inherited volume form  $\omega_S$  on  $S$  is defined by the property  $\nu^* \wedge \omega_S = \omega_{\mathbb{R}^3}$  at all points of  $S$ , where  $\nu^*$  is dual form to the oriented (outward) unit normal to  $S$ .

All forms of degree 2 are multiples of  $\omega_S$  by functions. In particular,

$$i_X(\nu^* \wedge \omega_S) = \nu \cdot X \omega_S,$$

i.e. we have to integrate the scalar product of the vector field  $X$  with the unit normal vector with respect to the standard volume on  $S$ . Thus, we have proved the following result formulated in the classical vector analysis style.

Actually, a simple check reveals that the above arguments work for all open submanifolds  $M \subset \mathbb{R}^n$  with boundary hypersurface  $S$  and vector fields  $X$ . The reader should easily verify this in detail.

#### DIVERGENCE THEOREM

**Theorem.** Let  $X$  be a vector field on a  $n$ -dimensional manifold  $M \subset \mathbb{R}^n$  with hypersurface boundary  $S$ . Then

$$\int_M \operatorname{div} X \, dx_1 \dots dx_n = \int_S X \cdot \nu \, dS,$$

where  $\nu$  is the oriented (outward) unit normal to  $S$  and  $dS$  stays for the volume inherited from  $\mathbb{R}^n$  on  $S$ .

Notice the 2-dimensional case coincides with the Green's theorem above.

**9.1.15. The original Stokes theorem.** If  $\omega$  is any linear form, then the integral of  $d\omega$  over a surface depends on the boundary curve only. This is the most classical *Stokes' theorem*:

#### THE CLASSICAL STOKES' THEOREM

In the case  $n = 3, k = 1$  we deal with a surface  $M$  in  $\mathbb{R}^3$  bounded by a curve  $C$ . The general linear forms are  $\omega = f \, dx + g \, dy + h \, dz$ , with the integral

$$\int_C f \, dx + g \, dy + h \, dz = \int_M d\omega,$$

where  $d\omega = \left( \frac{\partial h}{\partial y} - \frac{\partial g}{\partial z} \right) dy \wedge dz + \left( \frac{\partial f}{\partial z} - \frac{\partial h}{\partial x} \right) dz \wedge dx + \left( \frac{\partial g}{\partial x} - \frac{\partial f}{\partial y} \right) dx \wedge dy$ .

Again, we use the standard scalar product to identify the vector field  $X = f \frac{\partial}{\partial x} + g \frac{\partial}{\partial y} + h \frac{\partial}{\partial z}$  with the form  $\omega_X = f \, dx + g \, dy + h \, dz$ . Finally, reverting the above relation between the vector fields and two forms on  $\mathbb{R}^3$ , the 2-form  $d\omega_X$  can be identified with the vector field  $\operatorname{rot} X$ ,

$$d\omega_X = \omega_{\mathbb{R}^3}(\operatorname{rot} X, \cdot, \cdot).$$

**9.D.39.** Using d'Alembert's formula solve

$$u_{tt}(t, x) = u_{xx}(t, x) + \cos x, \quad t \in (0, \infty), \quad x \in \mathbf{R},$$

$$u(0, x) = x^2, \quad u_t(0, x) = \arctg x, \quad x \in \mathbf{R}.$$

**Solution.**

$$u_1(t, x) = \frac{\varphi(x - at) + \varphi(x + at)}{2} =$$

$$= \frac{(x - t)^2 + (x + t)^2}{2} = x^2 + t^2,$$

$$u_2(t, x) = \frac{1}{2a} \int_{x-at}^{x+at} \psi(\xi) d\xi = \frac{1}{2} \int_{x-t}^{x+t} \arctg \xi d\xi =$$

$$\left[ \xi \arctg \xi - \frac{1}{2} \ln(1 + \xi^2) \right]_{x-t}^{x+t} = \frac{1}{2} [(x + t) \arctg(x + t) -$$

$$-(x - t) \arctg(x - t) + \ln \sqrt{\frac{1 + (x - t)^2}{1 + (x + t)^2}}],$$

$$u_3(t, x) = \frac{1}{2a} \int_0^t \left( \int_{x-a(t-\sigma)}^{x+a(t-\sigma)} f(\sigma, \xi) d\xi \right) d\sigma =$$

$$= \frac{1}{2} \int_0^t \left( \int_{x-t+\sigma}^{x+t-\sigma} \cos \xi d\xi \right) d\sigma = \frac{1}{2} \int_0^t [\sin \xi]_{x-t+\sigma}^{x+t-\sigma} d\sigma =$$

$$\frac{1}{2} \int_0^t [\sin(x + t - \sigma) - \sin(x - t + \sigma)] d\sigma =$$

$$= \frac{1}{2} [\cos(x + t - \sigma)]_0^t + \frac{1}{2} [\cos(x - t + \sigma)]_0^t =$$

$$= \frac{1}{2} [-\cos(x + t) - \cos(x - t)] + \cos x.$$

$$u(t, x) = u_1(t, x) + u_2(t, x) + u_3(t, x).$$

**9.D.40.** Solve ( $x \in \mathbf{R}, t \in (0, \infty)$ )

- a)  $u_{tt} - u_{xx} = 0, u(0, x) = \sin x, u_t(0, x) = x \cos x$
- b)  $u_{tt} - u_{xx} = 0, u(0, x) = 2x, u_t(0, x) = \ln(1 + x^2)$
- c)  $u_{tt} - u_{xx} = \sin x, u(0, x) = x, u_t(0, x) = \frac{1}{x}$

**9.D.41.** Solve

$$u_{tt}(t, x) = u_{xx}(t, x) + 2 \sin x, \quad t \in (0, \infty), \quad x \in \mathbf{R},$$

$$u(0, x) = x^2 + 2x, \quad u_t(0, x) = x \cos x, \quad x \in \mathbf{R}.$$

**9.D.42.** Solve

$$u_{tt}(t, x) = u_{xx}(t, x) + 2 \cos x, \quad t \in (0, \infty), \quad x \in \mathbf{R},$$

$$u(0, x) = 2x^2 - x, \quad u_t(0, x) = x \sin x, \quad x \in \mathbf{R}.$$

This field is called the *rotation* or *curl* of the vector field  $X$ . The Stokes' theorem now reads:

$$\int_C \omega_X = \int_M \text{rot } X.$$

Consequently, the fields  $X$  with the property  $\omega_X = dF$  for some function  $F$  (the fields of gradients of functions), have got the property  $\text{rot } X = 0$ . They are called *conservative* (or *potential*) vector fields.

**9.1.16. Brouwer's theorem.** Among many useful fixed-point theorems, the *Brouwer theorem* is particularly nice.<sup>4</sup> We present a special case here. In particular, our formulation clearly must hold true for any homeomorphic image of the domain  $K$ . In fact, the convexity ensures that there are no "holes" inside of  $K$ . For example rotating an annulus  $r \leq \|x\| \leq R$  in the plane clearly does not have any fixed point.



**BROUWER'S FIXED-POINT THEOREM**

**Theorem.** Let  $K \subset \mathbf{R}^n$  be a compact convex submanifold of dimension  $n$  with boundary  $\partial K$ . Then every continuous mapping  $f : K \rightarrow K$  has got at least one fixed point  $x$ , i.e.  $f(x) = x$ .

**PROOF.** We shall restrict ourselves to the case of smooth manifold  $K$  and smooth mapping  $f$ . In fact, if a continuous  $f$  had no fixed point, then it is possible to approximate it by a smooth  $\tilde{f}$  without any fixed point (e.g., by taking a convolution  $\tilde{f} = f * \varphi$  with a suitable smooth kernel  $\varphi$  enjoying a very small support).

Assume  $f : K \rightarrow K$  is such a smooth mapping with  $f(x) \neq x$  for all  $x \in K$ . For each  $y \in K$ , consider the ray  $L$  from  $f(y)$  through  $y$ . This ray will leave  $K$  in the unique point  $F(y) \in \partial K$  and  $F : K \rightarrow \partial K$  is smooth. Notice we use the convexity assumption here (it is evident if  $K$  is a ball  $B$  of diameter  $r$  and the general case can be transformed to the ball case by "smoothly expanding"  $K$  from an inner point to a big enough ball  $B$ ). In particular, the construction implies  $F|_{\partial K} = \text{id}_{\partial K}$ . Thus, by our assumptions,  $F : K \rightarrow \partial K$  is a smooth retraction of  $K$  to its boundary.

Now, we may consider a smooth exterior form  $\omega \in \Omega^{n-1}(K)$  providing the standard (inherited) volume on  $\partial K$ , and we employ the general Stokes' theorem:

$$0 < \int_{\partial K} \omega = \int_{\partial K} F^* \omega = \int_K d(F^* \omega)$$

$$= \int_K F^*(d\omega) = \int_K F^*(0) = 0.$$

This is a contradiction and the theorem is proved. □

<sup>4</sup>Its 2-dimensional version is attributed to Luitzen Egbertus Jan Brouwer (1881-1966), a Dutch mathematician and philosopher, who should had noticed that when stirring a cup of coffee with sugar, there always stays at least one point fixed.

9.D.43.

$$u_{tt}(t, x) = 4u_{xx}(t, x) + \cos^2 x, \quad t \in (0, \infty), x \in \mathbf{R},$$

$$u(0, x) = x^3, \quad u_t(0, x) = \frac{1}{\sqrt{1+x^2}}, \quad x \in \mathbf{R}.$$

9.D.44.

$$u_{tt} = 9u_{xx} - 4 \cos x, \quad t \in (0, \infty), x \in \mathbf{R},$$

$$u(0, x) = x^2, \quad u_t(0, x) = xe^x, \quad x \in \mathbf{R}.$$

9.D.45. Solve parabolic equation

$$u_t = ku_{xx}, \quad t \in (0, \infty), x \in (0, l)$$

with *Dirichlet* boundary conditions

$$u(t, 0) = u(t, l) = 0, \quad t \in (0, \infty)$$

and initial condition

$$u(0, x) = \varphi(x) = x(l - x), \quad x \in (0, l).$$

**Solution.** Suppose that solution is in the form

$$u(t, x) = X(x) \cdot T(t)$$

and compute

$$u_t = XT', \quad u_{xx} = X''T,$$

dividing both sides of equation by  $XT$  we get

$$\frac{T'}{kT} = \frac{X''}{X} = -\lambda,$$

where  $\lambda$  is constant number, because left side depend only on variable  $t$  and right side only on variable  $x$ . Solving both ordinary equations

$$T' + \lambda kT = 0, \quad X'' + \lambda X = 0$$

we get

$$T(t) = Ce^{-\lambda kt}, \quad X(x) = C_1 e^{\mu x} + C_2 e^{-\mu x},$$

where  $\mu = \sqrt{-\lambda}$ . Using boundary conditions

$$X(0) = C_1 + C_2 = 0 \implies C_2 = -C_1$$

and

$$X(l) = C_1 (e^{\mu l} - e^{-\mu l}) = 0 \implies e^{2\mu l} = 1 = e^{2\pi ni}.$$

Numbers  $\mu_n = i\frac{n\pi}{l}$  and  $\lambda_n = \frac{n^2\pi^2}{l^2}$ , where  $n = 0, 1, 2, \dots$ . We have

$$T_n = C_n e^{-\frac{n^2\pi^2}{l^2} kt},$$

$$X_n = D_n (e^{i\frac{n\pi}{l}x} - e^{-i\frac{n\pi}{l}x}) = 2iD_n \sin\left(\frac{n\pi}{l}x\right).$$

Solution of equation and boundary condition is series

$$u(t, x) = \sum_{n=0}^{\infty} T_n X_n = \sum_{n=1}^{\infty} K_n \sin\left(\frac{n\pi}{l}x\right) e^{-\frac{n^2\pi^2}{l^2} kt}.$$

9.1.17. **Another kind of integration.** As we have seen, solutions to ODEs are flows of vector fields. As a modification, we can prescribe one-dimensional linear subspaces  $L_x \subset T_x M$  at each point of a manifold  $M$  and look for unparameterized curves  $P$  tangent to them at all points. This is a coordinate-free version of the ODE theory. Indeed, locally we may always choose a vector field  $X$  generating the spaces  $L_x$  and in each coordinate patch, the flow of  $X$  will provide the parameterized one-dimensional submanifolds  $P \subset M$  tangent to  $L_x$  at all points. A change of coordinates or  $X$  will change the parameterizations, but not the curves  $P$ .



If we want to describe an  $n$ -dimensional submanifold  $N \subset M$ ,  $1 < n < M$ , in a similar way, we define the  $n$ -dimensional subspaces  $D_x \subset T_x M$  for all  $x \in M$  and seek for a submanifold  $N$  with  $T_y N = D_y$  at all  $y \in N$ .

#### INTEGRABILITY OF DISTRIBUTIONS

The union  $D \subset TM$  of individual linear subspaces  $D_x \subset T_x M$ ,  $x \in M$ , is called a *distribution*  $D$  on  $M$ . We say that the distribution is  $n$ -dimensional and smooth if each fixed point  $x$  allows for a neighborhood  $U$  and  $n$  linearly independent smooth vector fields  $X_1, \dots, X_n$  generating  $D_y$  at all  $y \in U$ . The distribution is called *integrable* if for each point  $x \in M$ , there is a submanifold  $N \subset M$  such that  $x \in N$  and  $T_y N = D_y$  for all  $y \in N$ .

Our goal is to give necessary and sufficient conditions for smooth distributions to be integrable. Clearly, the case of  $n = 1$  is trivial, since we already know that the conditions are empty – each such distribution is integrable.

The core idea is to use the so called flow box theorem for vector fields proved in 8.3.15 and to exploit the individual flows of the chosen generators  $X_1, \dots, X_n$  in order to “draw” new coordinates, in which the integral submanifold would appear as given by  $x_{n+1} = 0, \dots, x_m = 0$ . The problem we face is that the flows do not commute in general and thus our idea will not work.

9.1.18. **Lie bracket of vector fields.** Fortunately, the commutativity of the flows is captured by a simple differential operation.

Consider two vector fields on  $\mathbb{R}^m$ ,  $X = X_1(x)\frac{\partial}{\partial x_1} + \dots + X_m(x)\frac{\partial}{\partial x_m}$ ,  $Y = Y_1(x)\frac{\partial}{\partial x_1} + \dots + Y_m(x)\frac{\partial}{\partial x_m}$ . The commutator of the derivatives of functions in the directions of these vector fields is

$$Y(Xf) - X(Yf) = \sum_{i,j} Y_i \frac{\partial}{\partial x_i} (X_j \frac{\partial f}{\partial x_j}) - X_i \frac{\partial}{\partial x_i} (Y_j \frac{\partial f}{\partial x_j})$$

$$= \sum_{i,j} (Y_j \frac{\partial X_i}{\partial x_j} - X_j \frac{\partial Y_i}{\partial x_j}) \frac{\partial f}{\partial x_i},$$

thanks to the commutativity of the second derivatives of  $f$ . Thus, the commutator of the two vector fields behaves as the

Coefficients  $K_n$  we get from initial condition

$$u(0, x) = \sum_{n=1}^{\infty} K_n \sin\left(\frac{n\pi}{l}x\right) = \varphi(x) = x(l-x).$$

So  $K_n$  are Fourier coefficient of odd extension of function  $\varphi(x)$  to the interval  $[-l, l]$ :

$$K_n = \frac{2}{l} \int_0^l x(l-x) \sin\left(\frac{n\pi}{l}x\right) dx.$$

After integration per partes we get  $K_n = \frac{4l^2}{n^3\pi^3} [(-1)^{n+1} + 1]$  and final solution

$$u(t, x) = \sum_{n=0}^{\infty} \frac{8l^2}{(2n+1)^3\pi^3} \sin\left(\frac{(2n+1)\pi}{l}x\right) e^{-\frac{(2n+1)^2\pi^2}{l^2}kt}.$$

**9.D.46.** Solve elliptic equation on square

$$\Delta u = 0, \quad x \in (0, 1), \quad y \in (0, 1)$$

with *Dirichlet* boundary conditions for  $x$

$$u(0, y) = u(1, y) = 0, \quad y \in (0, 1)$$

and boundary condition for  $y$

$$u(x, 0) = \varphi(x) = \sin(\pi x) \cos(\pi x), \quad x \in (0, 1),$$

$$u(x, 1) = \psi(x) = -\sin(\pi x), \quad x \in (0, 1).$$

**Solution.** Suppose again that solution is in the form

$$u(x, y) = X(x) \cdot Y(y) \implies \Delta u = X''Y + XY'' = 0.$$

Dividing equation by  $XY$  we get

$$\frac{X''}{X} = -\frac{Y''}{Y} = -\lambda.$$

Solution (using Dirichlet boundary conditions for  $x$ ) is

$$X_n = C_n \sin(n\pi x), \quad Y_n = K_n e^{n\pi y} + H_n e^{-n\pi y}, \quad n \in \mathbf{N}$$

and

$$u(x, y) = \sum_{n=1}^{\infty} X_n Y_n = \sum_{n=1}^{\infty} [k_n e^{n\pi y} + h_n e^{-n\pi y}] \sin(n\pi x).$$

Using boundary conditions for  $y$  we get

$$u(x, 0) = \sum_{n=1}^{\infty} (k_n + h_n) \sin(n\pi x) = \varphi(x),$$

$$u(x, 1) = \sum_{n=1}^{\infty} (k_n e^{n\pi} + h_n e^{-n\pi}) \sin(n\pi x) = \psi(x).$$

Numbers  $b_n = k_n + h_n$  are Fourier coefficients of odd extension of function  $\varphi(x)$  on  $[-1, 1]$ ,  $b_2 = \frac{1}{2}$  and  $b_n = 0$  for  $n \neq 2$ . Numbers  $\beta_n = k_n e^{n\pi} + h_n e^{-n\pi}$  are Fourier coefficients of odd extension of function  $\psi(x)$  on  $[-1, 1]$ ,  $\beta_1 = -1$  and  $\beta_n = 0$  for  $n \neq 1$ . We need to solve algebraic equations

$$k_2 + h_2 = \frac{1}{2}, \quad k_2 e^{2\pi} + h_2 e^{-2\pi} = 0,$$

$$k_1 + h_1 = 0, \quad k_1 e^{\pi} + h_1 e^{-\pi} = -1.$$

vector field  $[X, Y]$ ,

$$[X, Y] = \sum_{i,j=1}^m (Y_j \frac{\partial X_i}{\partial x_j} - X_j \frac{\partial Y_i}{\partial x_j}) \frac{\partial}{\partial x_i}.$$

This vector field is called the *Lie bracket*<sup>5</sup> of its arguments. It is easy to see that  $[ , ]$  is a bilinear antisymmetric operation (over the real scalars) on the differentiable vector fields, and expanding the commutators explicitly we arrive at the so called *Jacobi identity*

$$[X, [Y, Z]] = [[X, Y], Z] + [Y, [X, Z]]$$

valid for all triples of vector fields, and the *Leibnitz derivative property*

$$[X, fY] = (Xf)Y + f[X, Y].$$

□

**Remark.** In fact, it is quite straightforward to see that the vector fields  $X$  and the diffeomorphisms  $\text{Fl}_t^X$  in their flows are linked in a very similar manner to the square matrices  $A$  and their exponential images  $e^{tA}$ . The Lie bracket encodes the composition of the diffeomorphisms like the commutators of matrices encode the matrix multiplication. Thus, it is not surprising that the flows of two vector fields are commuting if and only if their Lie bracket vanishes. We shall not go into the technical proof here since we shall not need the result explicitly below.

**9.1.19. Back to distributions.** We say that  $D \subset TM$  is an *involutive distribution* if for all vector fields  $X, Y$  valued in  $D$ , their Lie bracket  $[X, Y]$  has got values in  $D$ , too.

FROBENIUS' THEOREM

**Theorem.** Let  $D \subset TM$  be a smooth  $n$ -dimensional distribution in an  $m$ -dimensional manifold  $M$ . Then  $D$  is integrable if and only if it is involutive.

**PROOF.** Remind integrability means the local existence of the integral submanifolds through each point in  $M$ . One of the implications of the proof is nearly trivial.

If  $D$  is integrable, then through each  $x \in M$  there is the integral submanifold  $N$ . Consider the embedding  $i : N \rightarrow M$  and any vector fields  $\tilde{X}, \tilde{Y}$  on  $M$  valued in  $D$ . Since  $D_y = T_y N$  for all  $y \in N$ ,  $\tilde{X}$  and  $\tilde{Y}$  are tangent to  $i(N) \subset M$ . We claim that the restriction of the Lie bracket  $[\tilde{X}, \tilde{Y}]$  to  $i(N)$  is the image  $i_*([X, Y])$ , where the vector fields  $X, Y$  are viewed as the given fields on  $N$ , i.e.,  $i_*X(x) = \tilde{X}(i(x))$ ,  $i_*Y(x) = \tilde{Y}(i(x))$ . Thus, the bracket has to be in the image again. The latter claim is a consequence of a more general statement:

<sup>5</sup>Marius Sophus Lie (1842–1899) was an excellent Norwegian mathematician, the father of the Lie theory. Originally invented to deal with systems of partial differential equation via continuous groups of their symmetries, the theory of Lie groups and Lie algebras is nowadays in the core of a vast part of Mathematics. It is a pity we do not have time and space to devote more attention to this marvelous mathematical story in this textbook.

Finally

$$k_1 = \frac{1}{e^{-\pi} - e^{\pi}}, \quad h_1 = \frac{1}{e^{\pi} - e^{-\pi}},$$

$$k_2 = \frac{1}{2 - 2e^{4\pi}}, \quad h_2 = \frac{1}{2 - 2e^{-4\pi}},$$

other coefficients are all zero. Solution is

$$u(x, y) = \sin(\pi x) \left[ \frac{1}{e^{-\pi} - e^{\pi}} e^{\pi y} + \frac{1}{e^{\pi} - e^{-\pi}} e^{-\pi y} \right] +$$

$$+ \sin(2\pi x) \left[ \frac{1}{2 - 2e^{4\pi}} e^{2\pi y} + \frac{1}{2 - 2e^{-4\pi}} e^{-2\pi y} \right].$$

Make a test.  $\square$

**9.D.47.** Solve hyperbolic equation

$$u_{tt} = a^2 u_{xx}, \quad t \in (0, \infty), \quad x \in (0, \pi)$$

with Neumann boundary conditions

$$u_x(t, 0) = u_x(t, \pi) = 0, \quad t \in (0, \infty)$$

and initial conditions

$$u(0, x) = \varphi(x) = 0, \quad u_t(0, x) = 5 \cos(3x), \quad x \in (0, \pi).$$

**Solution.** Again, looking for solution in the form

$$u(t, x) = X(x) \cdot T(t)$$

we get

$$XT'' = a^2 X''T \implies \frac{T''}{a^2 T} = \frac{X''}{X} = -\lambda.$$

So  $X(x) = C_1 e^{\mu x} + C_2 e^{-\mu x}$ , where  $\mu = \sqrt{-\lambda}$  and using Neumann boundary condition we have

$$X'(0) = \mu(C_1 - C_2) = 0 \implies C_1 = C_2,$$

$$X'(\pi) = \mu C_1 (e^{\mu\pi} - e^{-\mu\pi}) = 0,$$

$$e^{2\mu\pi} = 1 = e^{i2n\pi} \implies \mu_n = in, \lambda_n = -n^2, n = 0, 1, 2, \dots$$

We have solutions

$$X_n(x) = C_1 (e^{inx} + e^{-inx}) = 2C_1 \cos(nx),$$

$$T_n(t) = a_n \cos(nat) + b_n \sin(nat),$$

$$u(t, x) = \sum_{n=0}^{\infty} X_n T_n =$$

$$= \sum_{n=0}^{\infty} [A_n \cos(nat) + B_n \sin(nat)] \cos(nx).$$

Now we have to apply initial conditions

$$u(0, x) = \sum_{n=0}^{\infty} A_n \cos(nx) = \varphi(x),$$

$$u_t(0, x) = \sum_{n=0}^{\infty} naB_n \cos(nx) = \psi(x)$$

and numbers  $A_n$  are Fourier coefficients of even extension of function  $\varphi(x)$  to  $[-\pi, \pi]$ :

$$A_0 = \frac{1}{\pi} \int_0^{\pi} \varphi(x) dx, \quad A_n = \frac{2}{\pi} \int_0^{\pi} \varphi(x) \cos(nx) dx,$$

**Claim.** If  $\varphi : N \rightarrow M$  is a smooth map and two couples of vector fields  $X, Y$  and  $\tilde{X}, \tilde{Y}$  satisfy

$$T\varphi \circ X = \tilde{X} \circ \varphi, \quad T\varphi \circ Y = \tilde{Y} \circ \varphi,$$

then their Lie brackets satisfy the same relation:

$$T\varphi \circ [X, Y] = [\tilde{X}, \tilde{Y}] \circ \varphi.$$

Indeed, consider a smooth function  $f$  on  $M$  and compute, using  $X(f \circ \varphi)(x) = (T\varphi X)f = (\tilde{X} \circ \varphi)(x)f = \tilde{X}(\varphi(x))f$ , and similarly for  $Y$ :

$$\begin{aligned} [X, Y](f \circ \varphi)(x) &= XY(f \circ \varphi)(x) - YX(f \circ \varphi)(x) \\ &= X((\tilde{Y}f) \circ \varphi)(x) - Y((\tilde{X}f) \circ \varphi)(x) \\ &= \tilde{X}(\tilde{Y}f)(\varphi(x)) - \tilde{Y}(\tilde{X}f)(\varphi(x)) \\ &= ([\tilde{X}, \tilde{Y}]f) \circ \varphi(x). \end{aligned}$$

Now we employ the latter claim for the inclusion  $i$  in the role of  $\varphi$  and obviously every integrable distribution must be involutive.



As we already revealed, each one-dimensional distribution is involutive and locally integrable. The main idea of the proof is to start with any set of (locally) generating vector fields for  $D$ , to use some nice coordinates with respect to the first vector field and to employ the induction on the dimension to the rest of them.

Assume the theorem is true for dimensions less than  $n$  and consider an involutive smooth distribution  $D$  of dimension  $n$ , generated by fields  $X_1, \dots, X_n$ . Actually, we shall prove a much stronger version of the theorem. We claim that if  $D$  is involutive, then there are coordinates  $(x_1, \dots, x_m)$  around each point  $x \in M$ , such that the equations  $x_{n+1} = a_{n+1}, \dots, x_m = a_m$  with small constants  $a_i$  are defining all the individual integral submanifolds of  $D$  through points close to  $x$ . This is indeed true in dimension  $n = 1$ .

By the flowbox theorem 8.3.15, for each point  $x \in M$  there are coordinate functions  $y_1, \dots, y_m$  on a neighborhood  $U$  of  $x$ , for which  $X_1 = \frac{\partial}{\partial y_1}$ . Let us consider the submanifold  $Q \subset M$  defined by  $y_2 = 0, \dots, y_m = 0$  and the “projections”  $Y_j$  of the other fields to make them tangent to  $Q$ . This requires that  $Y_j$  leave constant the coordinate  $y_1$ , i.e. we set

$$Y_j = X_j - X_j(y_1)X_1, \quad j = 2, \dots, m.$$

Indeed, this definition ensures  $Y_j(y_1) = 0$  and thus the fields are tangent to  $Q$  as required. We leave  $Y_1 = X_1$ , and clearly  $Y_1, \dots, Y_n$  generate the same involutive distribution  $D$ . Thus

$$[Y_i, Y_j] = \sum_{i,j} c_{ijk} Y_k$$

for some set of functions  $c_{ijk}$ . Moreover, we may view  $Q$  as one leaf of all the subsets defined by  $y_2 = b_2, \dots, y_m = b_m$  with small constants  $b_i$  and there is the projection  $p : U \rightarrow Q$  forgetting the first coordinate.

On the submanifold  $Q$ , there is the  $(n-1)$ -dimensional involutive distribution  $\tilde{D}$  generated by the fields  $\tilde{Y}_i = Y_i|_Q$ ,  $i = 2, \dots, n$  (notice we again use the argument from the

numbers  $naB_n$  are Fourier coefficients of even extension of function  $\psi(x)$  to  $[-\pi, \pi]$ :

$$B_0 = \frac{1}{na\pi} \int_0^\pi \psi(x) dx, \quad B_n = \frac{2}{na\pi} \int_0^\pi \psi(x) \cos(nx) dx.$$

Because  $\varphi(x) = 0$ , all  $A_n$  are zero,  $\psi(x) = 5 \cos(3x)$ , so  $B_3 = \frac{5}{3a}$  and all other  $B_n$  are again zero. Finally solution of our problem is

$$u(t, x) = \frac{5}{3a} \sin(3at) \cos(3x).$$

Make a test.

**9.D.48.** Solve

$$\begin{aligned} \Delta u &= u_{xx} + u_{yy} = 0, \quad (x, y) \in (0, \pi) \times (0, \pi), \\ u(x, 0) &= u(x, \pi) = 0, \quad x \in (0, \pi) \\ u(\pi, y) &= 0, \quad u(0, y) = \sin y, \quad y \in (0, \pi). \end{aligned}$$

**9.D.49.** Solve

$$\begin{aligned} u_{xx} &= u_{tt}, \quad x \in (0, l), \quad t \in (0, \infty), \\ u_x(t, 0) &= u_x(t, l) = 0, \quad t \in (0, \infty), \\ u(0, x) &= -\cos \frac{\pi x}{l}, \quad x \in (0, l), \\ u_t(0, x) &= \cos^2 \frac{\pi x}{l} - \sin^2 \frac{\pi x}{l}, \quad x \in (0, l). \end{aligned}$$

**9.D.50.** Solve

$$\begin{aligned} u_{xx} &= u_t, \quad x \in (0, \pi), \quad t \in (0, \infty), \\ u(t, 0) &= u(t, \pi) = 0, \quad t \in (0, \infty), \\ u(0, x) &= 2 \cos(3x) \sin x, \quad x \in (0, \pi). \end{aligned}$$

**9.D.51.** Solve

$$\begin{aligned} u_{xx} &= u_{tt}, \quad x \in (0, \pi), \quad t \in (0, \infty), \\ u(t, 0) &= u(t, \pi) = 0, \quad t \in (0, \infty), \\ u(0, x) &= -6 \sin(2x) \cos(2x), \quad x \in (0, \pi), \\ u_t(0, x) &= -\sin x + 4 \sin x \cos x, \quad x \in (0, \pi). \end{aligned}$$

**9.D.52.** Solve

$$\begin{aligned} u_t &= u_{xx}, \quad x \in (0, a), \quad t \in (0, \infty), \\ u(t, 0) &= u_x(t, a) = 0, \quad t \in (0, \infty), \\ u(0, x) &= x(2a - x), \quad x \in (0, a). \end{aligned}$$

**9.D.53.**

$$\begin{aligned} u_{tt} &= u_{xx}, \quad x \in (0, A), \quad t \in (0, \infty), \\ u(t, 0) &= u(t, A) = 0, \quad t \in (0, \infty), \\ u(0, x) &= x(A - x), \quad u_t(0, x) = 0, \quad x \in (0, A). \end{aligned}$$

beginning of the proof about the brackets of restricted fields). Now, our assumption says we find suitable coordinates  $(q_2, \dots, q_m)$  on  $Q$  around the point  $x \in Q$ , so that for all small constants  $b_{n+1}, \dots, b_m$ , the integral submanifolds of  $\tilde{D}$  are defined by  $q_{n+1} = b_{n+1}, \dots, q_m = b_m$ .

Finally, we need to adjust the original coordinate functions  $y_i$  all over the neighborhood  $U$  of  $x$ . The obvious idea is to use the flow of  $X_1 = Y_1$  to extend the latter coordinates on  $Q$ . Thus we define the coordinate functions in all  $y \in U$  using the projection  $p$ ,

$$x_1(y) = y_1(y), x_2(y) = q_2(p(y)), \dots, x_m = q_m(p(y)).$$

The hope is that all submanifolds  $N$  given by equations  $x_{n+1} = b_{n+1}, \dots, x_m = b_m$  (for small  $b_j$ ) will be tangent to all fields  $Y_1, \dots, Y_n$ . Technically, this means  $Y_i(x_j) = 0$  for all  $i = 1, \dots, n, j = n + 1, \dots, m$ . By our definition, this is obvious for the restriction to  $Q$ , and obviously  $Y_1(x_j) = 0$  in all other points, too.

Let us look closely on what is happening with one of our functions  $Y_i(x_j)$  along the flows of the field  $X_1$ . We easily compute with the help of the definition of the Lie bracket

$$\begin{aligned} \frac{\partial}{\partial x_1}(Y_i(x_j)) &= Y_1(Y_i(x_j)) = Y_i(Y_1(x_j)) + [Y_1, Y_i](x_j) \\ &= Y_i(Y_1(x_j)) + c_{1i1}Y_1(x_j) + \sum_{k=2}^m c_{1ik}Y_k(x_j) \\ &= \sum_{k=2}^m c_{1ik}Y_k(x_j). \end{aligned}$$

This is a system of linear ODEs for the unknown functions  $Y_i(x_j)$  in one variable  $x_1$  along the flow lines of  $Y_1$ . The initial condition at the point in  $Q$  is zero and thus this constant zero value has to propagate along the flow lines, as requested.

The induction step is complete.  $\square$

**9.1.20. Formulation via exterior forms.** As we know from linear algebra, a vector subspace of codimension  $k$  is defined by  $k$  independent linear forms. Thus, every smooth  $n$ -dimensional distribution  $D \subset TM$  on a manifold  $M$  can be (at least) locally defined by  $m - n$  linear forms  $\omega_j$  on  $M$ .

A direct computation in coordinates reveals that the differential of linear form  $\omega$  evaluates on two vector fields as follows

$$(1) \quad d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y]).$$

Indeed, if  $X = \sum_i X_i \frac{\partial}{\partial x_i}$ ,  $Y = \sum_i Y_i \frac{\partial}{\partial x_i}$ ,  $\omega = \sum_i \omega_i dx_i$ , then

$$\begin{aligned} X(\omega(Y)) - Y(\omega(X)) &= \sum_{i,j} (X_i \frac{\partial}{\partial x_i} (\omega_j Y_j) - Y_i \frac{\partial}{\partial x_i} (\omega_j X_j)) \\ &= \sum_{i,j} (X_i \frac{\partial \omega_j}{\partial x_i} Y_j - Y_i \frac{\partial \omega_j}{\partial x_i} X_j + \omega_j (X_i \frac{\partial Y_j}{\partial x_i} - Y_i \frac{\partial X_j}{\partial x_i})) \\ &= d\omega(X, Y) + \omega([X, Y]). \end{aligned}$$

Thus, the involutivity of a distribution defined by linear forms  $\omega_{n+1}, \dots, \omega_m$  should be closely linked to properties of the

9.D.54.

$$\begin{aligned} \Delta u &= u_{xx} + u_{yy} = 0, \quad x \in (0, a), \quad y \in (0, a), \\ u(0, y) &= u(a, y), \quad y \in (0, a), \\ u(x, a) &= 0, \quad u(x, 0) = x(a - x), \quad x \in (0, a). \end{aligned}$$

9.D.55.

$$\begin{aligned} u_t &= 4u_{xx}, \quad x \in (0, 1), \quad t \in (0, \infty), \\ u_x(t, 0) &= u(t, 1) = 0, \quad t \in (0, \infty), \\ u(0, x) &= x - 1, \quad x \in (0, 1). \end{aligned}$$

9.D.56. Find bounded solution of Laplace equation and boundary condition on the circle (*Dirichlet internal boundary problem*):

$$\begin{aligned} \Delta u(x, y) &= 0, \quad x^2 + y^2 < A^2, \\ u(A \cos \varphi, A \sin \varphi) &= \alpha(\varphi) = \sin^4 \varphi. \end{aligned}$$

**Solution.** After transformation to the polar coordinates

$$x = r \cos \varphi, \quad y = r \sin \varphi,$$

Laplace equation for function  $u = u(r, \varphi)$  become

$$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\varphi\varphi} = 0.$$

We are looking for solution in the form  $u(r, \varphi) = R(r) \cdot \Phi(\varphi)$ , where  $\Phi(\varphi + 2\pi) = \Phi(\varphi)$  and  $u(A, \varphi) = \alpha(\varphi) = \sin^4 \varphi$ ,  $u(0, \varphi) = C$  is constant. Let's compute

$$R''\Phi + \frac{1}{r}R'\Phi + \frac{1}{r^2}R\Phi'' = 0,$$

multiplying by  $r^2$  and dividing by  $R\Phi$  we get

$$r^2 \frac{R'}{R} + r \frac{R'}{R} = -\frac{\Phi'}{\Phi} = -\lambda.$$

Solving equation

$$\Phi'' - \lambda\Phi = 0 \implies \Phi = B_1 e^{\mu\varphi} + B_2 e^{-\mu\varphi},$$

where  $\mu = \sqrt{\lambda}$ , using condition  $\Phi(\varphi) = \Phi(\varphi + 2\pi)$ ,  $\Phi(\varphi)$  have to be periodical with period  $2\pi$ , so

$$\mu = in \quad \text{and} \quad \lambda_n = -n^2, \quad n \in \mathbf{N}_0.$$

Finally for function  $\Phi(\varphi)$  we have

$$\Phi_n = a_n \cos(n\varphi) + b_n \sin(n\varphi).$$

Let's solve equation for  $R = R(r)$ , which is Euler

$$r^2 R'' + rR' - n^2 R_n = 0.$$

Substituing  $s = \ln t$ ,  $r = e^t$ ,

$$R'_n = \frac{1}{r} \dot{R}_n, \quad R''_n = (R'_n)' = -\frac{1}{r^2} \dot{R}_n + \frac{1}{r^2} \ddot{R}_n,$$

we get

$$\begin{aligned} \ddot{R}_n - n^2 R_n &= 0, \\ R_n(t) &= \begin{cases} C_0 + D_0 t, & n = 0, \\ C_n e^{nt} + D_n e^{-nt}, & n \in \mathbf{N}, \end{cases} \end{aligned}$$

- differentials on the common kernel. Indeed, there is the following version of the latter theorem:

FROBENIUS' THEOREM

**Theorem.** *The distribution  $D$  defined on an  $m$ -dimensional manifold  $M$  by  $(m - n)$  independent smooth linear forms  $\omega_{n+1}, \dots, \omega_m$  is integrable if and only if there are linear forms  $\alpha_{ij}$  such that  $d\omega_k = \sum_{\ell} \alpha_{k\ell} \wedge \omega_{\ell}$ .*

- **PROOF.** Let us write  $\omega = (\omega_{n+1}, \dots, \omega_m)$  for the  $\mathbb{R}^{m-n}$ -valued form. The distribution is  $D = \ker \omega$ . Now, the formula (1) (applied to all components of  $\omega$ ) implies that involutivity of  $D$  is equivalent to  $d\omega|_{\ker \omega} = 0$ .

If the assumption of the theorem on the forms holds true,  $d\omega$  clearly vanishes on the kernel of  $\omega$  and therefore  $D$  is involutive, and one of the implications of the theorem is proved.

Next, assume  $D$  is integrable. By the stronger claim proved in the latter Frobenius theorem, for each point  $x \in M$ , there are coordinates  $(x_1, \dots, x_m)$  such that  $D$  is the common kernel of all  $dx_{n+1}, \dots, dx_m$ . In particular, our forms  $\omega_j$  are linear combinations (over functions) of the latter  $(m - n)$  differentials. Moreover, there must be smooth invertible matrices of functions  $A = (a_{k\ell})$  such that

$$dx_k = \sum_{\ell} a_{k\ell} \omega_{\ell}, \quad k, \ell = n + 1, \dots, m.$$

Finally,  $d\omega_k$  includes only terms with  $dx_i \wedge dx_j$  with  $j > n$  and all  $dx_j$  can be expressed via our forms  $\omega_{\ell}$  from the previous equation. Thus the differentials have got the requested forms. □

## 2. Remarks on Partial Differential Equations

The aim of our excursion into the landscape of differential equations is modest. We do not have space in this rather elementary guide to come close enough to this subtle, beautiful, and extremely useful part of mathematics dealing with differential equations. Still we mention a few issues.

First, the simplest method reducing the problem to already mastered ordinary differential equations is explained, based on the so called characteristics. Then we show more simple methods how to get some families of solutions.

Next, we present a more complicated theoretical approach dealing with formal solvability of even higher order systems of differential equations and its convergence – the famous Cauchy-Kovalevskaya theorem. This is the only instance of general existence and uniqueness theorem for differential equations involving partial derivatives. Unfortunately, it does not cover many of interesting problems of practical importance.

Finally, we display a few classical methods to solve boundary problems involving some of the most common equations of second order.



$$R_n(r) = \begin{cases} C_0 + D_0 \ln r, & n = 0, \\ C_n r^n + D_n r^{-n}, & n \in \mathbf{N}. \end{cases}$$

Solution has to be bounded for  $r \rightarrow 0$  (inside circle), so  $D_0 = D_n = 0$  and finally

$$\begin{aligned} u(r, \varphi) &= R_n(r) \cdot \Phi_n(\varphi) = \\ &= a_0 C_0 + \sum_{n=1}^{\infty} C_n r^n [a_n \cos(n\varphi) + b_n \sin(n\varphi)]. \end{aligned}$$

Coefficients we compute from boundary condition

$$U(A, \varphi) = K_0 + \sum_{n=1}^{\infty} A^n [K_n \cos(n\varphi) + H_n \sin(n\varphi)] =$$

$$\begin{aligned} &= \alpha(\varphi) = \sin^4 \varphi = \left( \frac{1 - \cos(2\varphi)}{2} \right)^2 = \\ &= \frac{1 - 2\cos(2\varphi) + \cos^2(2\varphi)}{4} = \\ &= \frac{1}{4} - \frac{\cos(2\varphi)}{2} + \frac{1}{4} \left( \frac{1 + \cos(4\varphi)}{2} \right) = \\ &= \frac{3}{8} - \frac{\cos(2\varphi)}{2} + \frac{1}{8} \cos(4\varphi), \end{aligned}$$

$K_0 = 3/8$ ,  $K_2 = \frac{-1}{2A^2}$ ,  $K_4 = \frac{1}{8A^4}$  and all other  $K_n, H_n$  are zero.

$$\begin{aligned} u(r, \varphi) &= \frac{3}{8} - \frac{r^2}{2A^2} \cos(2\varphi) + \frac{r^4}{8A^4} \cos(4\varphi), \\ u(x, y) &= \frac{3}{8} - \frac{x^2 - y^2}{2A^2} + \frac{x^4 + y^4 - 6x^2y^2}{8A^4}, \end{aligned}$$

where we use  $\cos(2\varphi) = \cos^2 \varphi - \sin^2 \varphi = \frac{x^2}{r^2} - \frac{y^2}{r^2}$ ,  $\cos(4\varphi) = \cos^2(2\varphi) - \sin^2(2\varphi)$ .  $\square$

**9.D.57.** Find bounded solution of Laplace equation and boundary condition on the circle (*Dirichlet external boundary problem*):

$$\begin{aligned} \Delta u(x, y) &= 0, \quad x^2 + y^2 > A^2, \\ u(A \cos \varphi, A \sin \varphi) &= \alpha(\varphi) = \cos^2 \varphi. \end{aligned}$$

**Solution.** From previous exercise we get

$$\begin{aligned} \Phi_n &= a_n \cos(n\varphi) + b_n \sin(n\varphi), \quad n \in \mathbf{N}_0, \\ R_n(r) &= \begin{cases} D_0 + C_0 \ln r, & n = 0, \\ C_n r^n + D_n r^{-n}, & n \in \mathbf{N}. \end{cases} \end{aligned}$$

Functions  $R_n(r)$  have to be bounded for  $r \rightarrow \infty$  (outside of the circle), so  $C_0 = C_n = 0$  and we have

$$\begin{aligned} u(r, \varphi) &= R_n(r) \cdot \Phi_n(\varphi) = \\ &= a_0 D_0 + \sum_{n=1}^{\infty} D_n r^{-n} [a_n \cos(n\varphi) + b_n \sin(n\varphi)]. \end{aligned}$$

Coefficients we compute from boundary condition

$$\begin{aligned} U(A, \varphi) &= K_0 + \sum_{n=1}^{\infty} A^{-n} [K_n \cos(n\varphi) + H_n \sin(n\varphi)] = \\ &= \alpha(\varphi) = \cos^2 \varphi = \frac{1 - \cos(2\varphi)}{2}. \end{aligned}$$

**9.2.1. Initial observations.** In practical problems, we often meet equations relating unknown functions of more variables and their derivatives. We already handled the very special case where the relations concerned functions  $x(t)$  of just one variable  $t$ . More explicitly, we dealt with vector equations



$$x^{(k)} = F(t, x, \dot{x}, \ddot{x}, \dots, x^{(k-1)}), \quad F: \mathbb{R}^{n \times k+1} \rightarrow \mathbb{R}^n,$$

where the dots over  $x \in \mathbb{R}^n$  meant the (iterated) derivatives of  $x(t) = (x_1(t), \dots, x_n(t))$ , up to the order  $k$ . The goal was to find a (vector) curve  $x(t)$  in  $\mathbb{R}^n$  which makes this equation valid.

Two more comments are due: 1) we can omit the explicit appearance of  $t$  on the cost of adding one more variable and equation  $\dot{x}_0 = 1$ ; and 2) giving new names to the iterated derivatives  $x_j = x^{(j)}$  and adding equations  $\dot{x}_j = x_{j+1}$ ,  $j = 1, \dots, k-1$ , we reduce always the problem to a first order system of equations (on a much bigger space).

Thus, we should like to work similarly with the equations

$$F(x, y, u_x, u_{xx}, u_{xy}, u_{yy}, \dots) = 0,$$

where  $u$  is an unknown function (possibly vector valued) of two variables  $x$  and  $y$  (or even more variables) and, as usual, the indices denote the partial derivatives. Even if we expect the implicit equation to be solved in some sense with respect to some of the highest partial derivatives, we cannot hope for a general existence and uniqueness result similar to the ODE case.

Let us start with a most simple example illustrating the general problem related to the choice of the initial conditions.

**9.2.2. The simplest linear case.** Consider one real function  $u = u(x, y)$ , subject to the linear homogeneous equation

$$(1) \quad a(x, y)u_x + b(x, y)u_y = 0$$

where  $a$  and  $b$  are known functions of two variables defined for  $x, y$  in a domain  $\Omega \subset \mathbb{R}^2$ . We consider the equation in the tubular domain  $\Omega \times \mathbb{R} \subset \mathbb{R}^3$ . Usually,  $\Omega$  is an open set together with a nice boundary, a curve  $\partial\Omega$  in our case.

An obvious simple idea suggests to write  $\Omega$  as a union of non-intersecting curves and look for  $u$  constant along those curves. Moreover, if those curves were transversal to the boundary  $\partial\Omega$ , then initial conditions along the boundary should extend inside of  $\Omega$ . Thus, consider such a potentially existing curve  $c(t) = (x(t), y(t))$  and write

$$0 = \frac{d}{dt} u(c(t)) = u_x(c(t))\dot{x}(t) + u_y(c(t))\dot{y}(t).$$

This yields the conditions for the requested curves:

$$(2) \quad \dot{x} = a(x, y), \quad \dot{y} = b(x, y).$$

Since  $u$  is considered constant along the curve, we obtain a unique possibility for the function  $u$  along the curves for all initial conditions  $x(0), y(0)$ , and  $u(x(0), y(0))$ , if the coefficients  $a$  and  $b$  are at least Lipschitz in  $x$  and  $y$ .

The latter curves are called the *characteristics* of the first order partial differential equation (1) and they are solutions

We get  $K_0 = \frac{1}{2}$ ,  $K_2 = \frac{A^2}{2}$  and all other  $K_n, H_n$  are zero.

$$u(r, \varphi) = \frac{1}{2} + \frac{1A^2}{2r^2} \cos(2\varphi),$$

$$u(x, y) = \frac{1}{2} + \frac{x^2 - y^2}{2(x^2 + y^2)^2} A^2.$$

**9.D.58.** Solve internal Dirichlet problem on the circle:

$$\Delta u(x, y) = 0, \quad x^2 + y^2 < A^2,$$

$$u(A \cos \varphi, A \sin \varphi) = \alpha(\varphi) = \sin \varphi \cos \varphi.$$

**9.D.59.** Solve external Dirichlet problem on the circle:

$$\Delta u(x, y) = 0, \quad x^2 + y^2 > A^2,$$

$$u(A \cos \varphi, A \sin \varphi) = \alpha(\varphi) = 2 \sin^2 \varphi + 3 \cos^2 \varphi.$$

**9.D.60.** Solve internal Dirichlet problem on the circle:

$$\Delta u(x, y) = 0, \quad x^2 + y^2 < 1,$$

$$u(\cos \varphi, \sin \varphi) = \alpha(\varphi) = \sin \varphi + \cos^2 \varphi.$$

**9.D.61.** Solve parabolic equation

$$u_t = \kappa u_{xx}, \quad x \in \mathbf{R}, \quad t \in (0, \infty),$$

with initial condition

$$u(0, x) = \varphi(x) = \begin{cases} 1 & \text{for } x \in (-1, 1), \\ 0 & \text{otherwise.} \end{cases}$$

Suppose  $u(t, x) \rightarrow 0$  and  $u_x(t, x) \rightarrow 0$  for  $x \rightarrow \pm\infty$ .

**Solution.** Let's compute Fourier image

$$\begin{aligned} \tilde{u}_{xx}(\xi) &= \int_{-\infty}^{\infty} u_{xx} e^{-i\xi x} dx = \left| \begin{array}{ll} a' = u_{xx} & a = u_x \\ b = e^{-i\xi x} & b' = -i\xi e^{-i\xi x} \end{array} \right| = \\ &= [u_x e^{-i\xi x}]_{-\infty}^{\infty} - (-i\xi) \int_{-\infty}^{\infty} u_x(x) e^{-i\xi x} dx = \\ &= \left| \begin{array}{ll} a' = u_x & a = u \\ b = e^{-i\xi x} & b' = -i\xi e^{-i\xi x} \end{array} \right| = [u_x e^{-i\xi x}]_{-\infty}^{\infty} + \\ &+ i\xi [u e^{-i\xi x}]_{-\infty}^{\infty} + (-i\xi)^2 \int_{-\infty}^{\infty} u e^{-i\xi x} dx = 0 + 0 - \xi^2 \tilde{u}, \end{aligned}$$

because  $u \rightarrow 0$  and  $u_x \rightarrow 0$  for  $x \rightarrow \pm\infty$ . After Fourier transformation the equation became

$$\tilde{u}_t(t, \xi) = -\kappa \xi^2 \tilde{u}(t, \xi)$$

and together with initial condition  $\tilde{u}(0, \xi) = \tilde{\varphi}(\xi)$ , where  $\tilde{\varphi}(\xi)$  is Fourier image of  $\varphi(x)$ , we have Fourier image of solution

$$\tilde{u}(t, \xi) = \tilde{\varphi}(\xi) e^{-\kappa \xi^2 t},$$

of its *characteristic equations* (2). If the coefficients are differentiable in all variables, then also the solution  $u$  will be differentiable for differentiable choices of initial conditions on a curve transversal to the characteristics and we might have solved the problem (1) locally. Still it might fail.

Let us look at the homogeneous linear problem

$$\square \quad (3) \quad yu_x - xu_y = 0, \quad u(x, 0) = x.$$

We saw already the solutions to the characteristic equations

$$\dot{x} = y, \quad \dot{y} = -x$$

and the characteristics are circles with centers in the origin,  $x(t) = R \sin t, y(t) = R \cos t$ . If we choose any even differentiable function  $\psi(x) = u(x, 0)$  for the initial conditions at points  $(x, 0)$ , we are lucky to see that the solution will work. But for odd functions, e.g. our choice  $\psi(x) = x$ , there will be no solution of our problem in any neighbourhood of the origin. Clearly, this failure is linked to the fact that the origin is a singular point for characteristic equations.

**9.2.3. The quasi-linear case.** The situation seems to get more tricky once we add a nontrivial right-hand value  $f(x, y, u)$  to the equation (1), i.e. we try to solve the problem (allowing  $a$  and  $b$  to depend on  $u$ )

$$\circ \quad (1) \quad a(x, y, u)u_x + b(x, y, u)u_y = f(x, y, u).$$

But in fact, the very same idea leads to characteristic equations on  $\mathbb{R}^3$ , writing  $z = u(x, y)$  for the unknown function along the characteristics. Geometrically, we seek for a vector field tangent to all graphs of solutions in the tubular domain  $\Omega \times \mathbb{R}$ . Remind  $z = u(x, y)$ , restricted to a curve in the graph, implies  $\dot{z} = u_x \dot{x} + u_y \dot{y}$ , and thus we may set  $\dot{z} = f(x, y, z), \dot{x} = a(x, y, z), \dot{y} = b(x, y, z)$  in order to get such a *characteristic vector field*.

#### CHARACTERISTIC EQUATIONS AND INTEGRALS

The *characteristic equations* of the equation (1) are

$$(2) \quad \dot{x} = a(x, y, z), \quad \dot{y} = b(x, y, z), \quad \dot{z} = f(x, y, z).$$

This autonomous system of three equations is uniquely solvable for each initial condition if  $a, b$ , and  $f$  are Lipschitz.

A function  $\psi$  on  $\Omega \times \mathbb{R}$  which is constant on each flow line of the characteristic vector field, i.e.,  $\psi(x(t), y(t), z(t)) = \text{const}$  for all solutions of (2), is called an *integral* of the equation (1). If  $\psi_z \neq 0$ , then the implicit function theorem guarantees the unique existence of the function  $z = u(x, y)$  satisfying the chosen initial conditions.

Check yourself that the latter functions  $u$  are solutions to our problem. This approach covers the homogeneous case as well, we just consider the autonomous characteristic equations with  $\dot{z} = 0$  added.

Let us come back to our simple equation 9.2.2(3) and choose  $f(x, y, u) = y$  for the right-hand side. The characteristic equations yield  $x = R \sin t, y = R \cos t$  as before, while  $\dot{z} = y = R \cos t$  and hence  $z = R \sin t + z(0)$ . Thus, we

(solve this separable equation with initial condition thoroughly). If Fourier image of solution is product of two functions, the original solution has to be convolution of inverse Fourier images of this functions:

$$u(t, x) = \int_{-\infty}^{\infty} \varphi(y) \frac{1}{2\sqrt{\pi\kappa t}} e^{-\frac{(x-y)^2}{4\kappa t}} dy.$$

We have used the fact, that inverse Fourier image of Gaussian function  $e^{-\kappa\xi^2 t}$  is again Gaussian function  $\frac{1}{2\sqrt{\pi\kappa t}} e^{-\frac{x^2}{4\kappa t}}$  and formula for convolution

$$f * g(x) = \int_{-\infty}^{\infty} f(y)g(x - y)dy.$$

We will express  $u(t, x)$  using the *error function*

$$\operatorname{erf}(x) = \frac{2}{\pi} \int_0^x e^{-a^2} da.$$

Take  $\varphi(x)$  from assignment and compute

$$\begin{aligned} u(t, x) &= \int_{-1}^1 \frac{1}{2\sqrt{\pi\kappa t}} e^{-\frac{(x-y)^2}{4\kappa t}} dy = \\ &= \left| \begin{array}{l} a = \frac{x-y}{2\sqrt{\kappa t}} \\ da = \frac{-1}{2\sqrt{\kappa t}} dy \end{array} \right| = \int_{\frac{x+1}{2\sqrt{\kappa t}}}^{\frac{x-1}{2\sqrt{\kappa t}}} \frac{-1}{\sqrt{\pi}} e^{-a^2} da = \\ &= \frac{1}{\sqrt{\pi}} \int_0^{\frac{x+1}{2\sqrt{\kappa t}}} e^{-a^2} da - \frac{1}{\sqrt{\pi}} \int_0^{\frac{x-1}{2\sqrt{\kappa t}}} e^{-a^2} da, \\ u(t, x) &= \frac{1}{2} \left[ \operatorname{erf} \left( \frac{x+1}{2\sqrt{\kappa t}} \right) - \operatorname{erf} \left( \frac{x-1}{2\sqrt{\kappa t}} \right) \right]. \end{aligned}$$

**9.D.62.** Solve parabolic equation

$$u_t = u_{xx}, \quad x \in \mathbf{R}, \quad t \in (0, \infty),$$

with initial condition

$$u(0, x) = \varphi(x) = \begin{cases} 2 & \text{for } x \in (0, 1), \\ 0 & \text{otherwise.} \end{cases}$$

Suppose  $u(t, x) \rightarrow 0$  and  $u_x(t, x) \rightarrow 0$  for  $x \rightarrow \pm\infty$  and express solution using error function. ○

**E. Variational Problems**

**F. Complex analytic functions**

**9.F.1.** Check that the mapping  $z \mapsto z^n$ ,  $n \in \mathbb{N}$ , defined on the entire  $\mathbb{C}$  has got the complex derivative  $z \mapsto nz^{n-1}$ .

**Solution.** We can proceed two ways. Either check the definition of the complex derivative directly, cf. ??, or use the detour via the explicit expression of the mapping  $f(x+iy) = (x+iy)^n$  in two real coordinates  $x, y$  and see that its derivative is complex.

may choose  $\psi(x, y, z) = z - x$  as an integral of the equation, and the solution  $u(x, y) = x + C$  with any constant  $C$ .

Notice, there will be plenty of solutions here since we may add any solution of the homogenous problem, i.e. all functions of the form

$$(3) \quad u(x, y) = h(x^2 + y^2)$$

with any differentiable function  $h$ . Thus, the general solution  $u(x, y) = x + h(x^2 + y^2)$  depends on one function of one variable (the above constant  $C$  is a special case of  $h$ ).

We may also conclude that for “reasonable” curves  $\partial\Omega \subset \mathbb{R}^2$  (those transversal to the circles centred at the origin and not containing the origin) and “reasonable” initial value  $u|_{\partial\Omega}$  (we have to watch the multiple intersection of the circles with  $\partial\Omega$ ) there will be (at least locally) a unique solution extending the initial values to an open neighborhood of  $\partial\Omega$ .

Of course, we may similarly use characteristics and integrals for any finite number of variables  $x = (x_1, \dots, x_n)$  and equations of the form

$$a_1(x, u) \frac{\partial u}{\partial x_1} + \dots + a_n(x, u) \frac{\partial u}{\partial x_n} = f(x, u)$$

with the unknown function  $u = u(x_1, \dots, x_n)$ . As we shall see later, typically we obtain generic solutions dependent on one function of  $n - 1$  variables, similarly to the above example.

**9.2.4. Systems of equations.** Let us look what happens if we add more equations. There are two quite different ways how to couple the equations.

We may seek for an unknown vector valued functions  $u = (u_1, \dots, u_m) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , subject to  $m$  equations

$$(1) \quad A_i(x, u) \cdot \nabla u_i = f_i(x, u), \quad i = 1, \dots, m,$$

where the left hand side means the scalar product of a vector valued function  $A_i : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^n$  and the gradient vector of the function  $u_i$ . Such systems behave similarly as the scalar ones and we shall come back to them later. □

The other option leads to the so called *overdetermined systems* of equations. Actually we shall not pay more attention to this case in the sequel and so the reader might jump to 9.2.6 if getting lost.

Consider a (scalar) function  $u$  on a domain in  $\Omega \subset \mathbb{R}^n$  and its gradient vector  $\nabla u$ . For each matrix  $A = (a_{ij})$  with  $m$  rows and  $n$  columns, with differentiable functions  $a_{ij}(x, u)$  on  $\Omega \times \mathbb{R}$ , and the right hand value function  $F(x, u) : \Omega \times \mathbb{R} \rightarrow \mathbb{R}^m$ , we can consider the system of equations

$$(2) \quad A(x, u) \cdot \nabla u = F(x, u).$$

Of course, in both cases, we have got  $m$  individual equations of the type from the previous paragraph and we could apply the same idea of characteristic vector fields for all of them. The problem consists in coupling of the equations and obtaining possibly inconsistent necessary conditions from the individual characteristic fields.

The first very simple possibility repeats the computation with polynomials in one real variable and is shown in 9.4.1.  $\square$

Let us look at the overdetermined case now. We can get most close to the situation with the ordinary differential equations if  $A$  is invertible and we move it to the right hand side, arriving at the system of equations

$$(3) \quad \nabla u = A^{-1}(x, u) \cdot F(x, u) = G(x, u).$$

The simplest non-trivial case consists of two equations in two variables:

$$u_x = f(x, y, u), \quad u_y = g(x, y, u).$$

Geometrically, we describe the graph of the solution as a surface in  $\mathbb{R}^3$  by prescribing its tangent plane through each point. An obvious condition for the existence of such  $u$  is obtained by differentiating the equations and employing the symmetry of the higher order partial derivatives, i.e. the condition  $u_{xy} = u_{yx}$ . Indeed,

$$u_{xy} = f_y + f_u g = g_x + g_u f = u_{yx},$$

where we substituted the original equations after applying the chain rule. We shall see in a moment that this condition is also sufficient for the existence of the solutions. Moreover, if the solutions exist, then they are determined by their values in one point, similarly to the ordinary differential equations.

**9.2.5. Frobenius' theorem again.** Similarly, we can deal with the gradient  $\nabla u$  of an  $m$ -dimensional vector valued function  $u$ . For example, if  $m = 2$  and  $n = 2$  we are describing the tangent planes to the two-dimensional graph of the solution  $u$  in  $\mathbb{R}^4$ . In general we face  $mn$  equations

$$(1) \quad \frac{\partial u_p}{\partial x_i} = F_{pi}(x, u), \quad i = 1, \dots, n, \quad p = 1, \dots, m.$$

The necessary conditions imposed by the symmetry of higher order derivatives then read

$$(2) \quad \frac{\partial^2 u_p}{\partial x_i \partial x_j} = \frac{\partial F_{pi}}{\partial x_j} + \sum_q \frac{\partial F_{pi}}{\partial u_q} F_{qj} = \frac{\partial F_{pj}}{\partial x_i} + \sum_q \frac{\partial F_{pj}}{\partial u_q} F_{qi}$$

for all  $i, j$  and  $p$ .

Let us reconsider our problem from the geometric point of view now. We are seeking for the graph of the mapping  $u : \mathbb{R}^n \rightarrow \mathbb{R}^m$ . The equations (1) describe the  $n$ -dimensional distribution  $D$  on  $\mathbb{R}^{m+n}$  and the graphs of possible solutions  $u = (u_1, \dots, u_m)$  are just the integral manifolds of  $D$ . The distribution  $D$  is clearly defined by the  $m$  linear forms

$$\omega_p = du_p - \sum_i F_{pi} dx_i, \quad p = 1, \dots, m,$$

while the vector fields generating the common kernel of all  $\omega_p$  can be chosen as

$$X_i = \frac{\partial}{\partial x_i} + \sum_p F_{pi} \frac{\partial}{\partial u_p}.$$

Now we compute differentials  $d\omega_p$  and evaluate them on the fields  $X_i$

$$\begin{aligned} -d\omega_p &= \sum_{i,j} \frac{\partial F_{pi}}{\partial x_j} dx_j \wedge dx_i + \sum_{i,q} \frac{\partial F_{pi}}{\partial u_q} du_q \wedge dx_i \\ &= \sum_{i,j} \left( \frac{\partial F_{pi}}{\partial x_j} + \sum_q \frac{\partial F_{pi}}{\partial u_q} F_{qj} \right) dx_j \wedge dx_i \\ -d\omega_p(X_j, X_i) &= \left( \frac{\partial F_{pi}}{\partial x_j} + \sum_q \frac{\partial F_{pi}}{\partial u_q} F_{qj} \right) \\ &\quad - \left( \frac{\partial F_{pj}}{\partial x_i} + \sum_q \frac{\partial F_{pj}}{\partial u_q} F_{qi} \right). \end{aligned}$$

Thus, vanishing of the differentials on the common kernel is equivalent to the necessary conditions deduced above, and the Frobenius theorem says that the latter conditions are sufficient, too. We have proved the following:

**Theorem.** *The system of equations (1) admits solutions if and only if the conditions (2) are satisfied. Then the solutions are determined uniquely locally around  $x \in \Omega$  by the initial conditions  $u(x) \in \mathbb{R}^m$ .*

**Remark.** The Frobenius' theory deals with the so called *overdetermined systems of PDEs*, i.e. we have got too many equations and this causes obstructions towards their integrability. Although the case in the last paragraph sounds very special, the actual use of the theory consists in considering differential consequences of a given system until we reach a point, where the special theorem applies and gives not only further obstructions but also the sufficient conditions.

**9.2.6. General solutions to PDE's.** In a moment, we shall deal with diverse boundary conditions for the solutions of PDEs. In most cases we shall be happy to have good families of simple "guessed" solutions which are not subject to any further conditions. We talk about *general solutions* in this context. Unlike the situation with ODEs, we should not hope to get a universal expression for all possible solutions this way (although we can come close to that in some cases, cf. 9.2.3(3)). Instead, we often try to find the right superpositions (i.e. linear combinations) or integrals built from suitable general solutions.



Let us look at the simplest linear second order equations in two variables, homogeneous with constant coefficients:

$$(1) \quad Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = 0$$

where  $A, B, C, D, E, F$  are real constants and at least one of  $A, B, C$  is non-zero.

Similarly to the method of characteristics, we try to reduce the problem to ODEs. Let us again assume solution in the form  $u = f(p)$ , where  $f$  is an unknown function of  $p$  and  $p(x, y)$  should be nice enough to get close to solutions. The necessary derivatives are  $u_x = f'p_x$ ,  $u_y = f'p_y$ ,  $u_{xx} = f''p_x p_x + f'p_{xx}$ ,  $u_{xy} = f''p_x p_y + f'p_{xy}$ ,  $u_{yy} =$

$f''p_y p_y + f'p_{yy}$ . Thus (1) becomes too complicated in general, but restricting to affine  $p(x, y) = \alpha x + \beta y$  with constants  $\alpha, \beta$ , we arrive at

$$(2) (A\alpha^2 + 2B\alpha\beta + C\beta^2)f'' + (D\alpha + E\beta)f' + Ff = 0.$$

This is a nice ODE as soon as we fix the values of  $\alpha$  and  $\beta$ . Let us look at several simple cases of special importance.

Assume  $D = E = F = 0$ ,  $A \neq 0$ . Then, after dividing by  $\alpha^2$ , we solve the equation  $(A + 2B\frac{\beta}{\alpha} + C\frac{\beta^2}{\alpha^2})f'' = 0$  and the right choice of the ratio  $\lambda = \beta/\alpha \neq 0$  kills the entire coefficient at  $f''$ . Thus, (2) will hold true for any (twice differentiable) function  $f$  and we arrive at the general solution  $u(x, y) = f(p(x, y))$ , with  $p(x, y) = x + \lambda y$ . Of course, the behavior will very much depend on the number of real roots of the quadratic equation

$$A + 2B\lambda + C\lambda^2 = 0.$$

**The wave equation.** Put  $A = 1$ ,  $C = -\frac{1}{c^2}$ ,  $B = 0$ , thus our equation is  $u_{xx} = \frac{1}{c^2}u_{yy}$ , the *wave equation* in dimension 1. Then the equation  $1 - \frac{1}{c^2}\lambda^2 = 0$  has got two real roots  $\lambda = \pm c$ , and we obtain  $p = x \pm cy$  leading to the general solution

$$u(x, y) = f(x - cy) + g(x + cy)$$

with two arbitrary twice differentiable functions of one variable  $f$  and  $g$ .

In Physics, the equation models one-dimensional wave development in the space parametrized by  $x$  while  $y$  stays for the time. Notice  $c$  corresponds to the speed of the wave  $u(x, 0) = f(x) + g(x)$  initiated in the time  $y = 0$ , and while the  $f$  part moves forwards, the other part moves backwards. Indeed, imagine  $u(x, y) = f(x - cy)$  describes the displacement of a string at point  $x$  in time  $y$ . This remains constant along the lines  $x - cy = \text{constant}$ . Thus, a stationary observer sees the initial displacement  $u(x, 0)$  moving along  $x$ -axis with the speed  $c$ .

In particular, we see that the initial condition along a line in the plane is not enough to determine the solution, unless we request the solution will move only in one of the possible directions (i.e. we posit either  $f$  or  $g$  to be zero).

**The Laplace equation.** Now we consider  $A = C = 1$ ,  $B = 0$ , i.e. the equation  $u_{xx} + u_{yy} = 0$ . This is the *Laplace equation* in two dimensions and its solutions are called *harmonic functions*.

Proceeding as before, we obtain two imaginary solutions to the equation  $\lambda^2 + 1 = 0$  and our method produces  $p = x \pm iy$ , a complex valued function instead of the expected real one. This looks ridiculous, but we could consider  $f$  to be a mapping  $f : \mathbb{C} \rightarrow \mathbb{C}$  viewed as a mapping on the complex plane. Remind that some of such mappings have got differentials  $D^1 f(p)$  which actually are multiplications by complex numbers at each point, cf. ???. This is in particular true for any polynomial or converging power series. We may request that this property holds true for all iterated derivatives of this kind. In general, we call such functions on  $\mathbb{C}$  holomorphic and we discuss them in the last part of this chapter. The reader is

advised to come back to this exposition on general solutions to Laplace equation after reading through the beginning of the part on complex analytic functions below, starting in 9.4.1.

Now, assuming  $f$  is holomorphic, we can repeat the above computation and arrive again at

$$(\lambda^2 + 1)f''(p) = 0$$

independently of the choice of  $f$  (here  $f'(p)$  means the complex number given by the differential  $D^1 f$ ,  $f''(p)$  is the iteration of this kind of derivative). Moreover, the derivatives of vector valued functions are computed for the components separately and thus both the real and the imaginary part of the general solution  $f(x + iy) + g(x - iy)$  will be real general solutions.

For example, consider  $f(p) = p^2$  leading to

$$u(x, y) = (x + iy)^2 = (x^2 - y^2) + i2xy$$

and simple check shows that both terms satisfy the equation separately. Notice the two solutions  $x^2 - y^2$  and  $xy$  provide the bases of the 2-dimensional vector space of harmonic homogeneous polynomials of degree two.

**The diffusion equation.** Next assume  $A = \kappa$ ,  $B = C = D = F = 0$ , and add the first order term with  $E = -1$ . This provides the equation

$$u_y = \kappa u_{xx},$$

the *diffusion equation* in dimension one.

Applying the same method again, we arrive at the ODE

$$\kappa\alpha^2 f'' - \beta f' = 0$$

which is easy to solve. We know the solutions are found in the form  $f(p) = e^{\nu p}$  with  $\nu$  satisfying the condition  $\kappa\alpha^2\nu^2 - \beta\nu = 0$ . The zero solution is not interesting, thus we are left with the general solution to our problem by substituting  $p(x, y) = \alpha x + \beta y$  and  $\nu = \frac{\beta}{\kappa\alpha^2}$ :

$$u(x, y) = f(p) = e^{\frac{1}{\kappa}(\frac{\beta}{\alpha}x + \frac{\beta^2}{\alpha^2}y)}.$$

Again, a simple check reveals that this is a solution. But it is not very "general" – it depends just on two scalars  $\alpha$  and  $\beta$ . We have to find much better ways how to find solutions of such equations.

**9.2.7. Nonhomogeneous equations.** As always with linear equations, the space of solutions to the homogeneous linear equations is a real vector space (or complex, if we deal with complex valued solutions).

Let us write the equation as  $Lu = 0$ , where  $L$  is the differential operator on the left hand side. For instance,

$$L = A \frac{\partial^2}{\partial x^2} + B \frac{\partial^2}{\partial x \partial y} + C \frac{\partial^2}{\partial y^2} + D \frac{\partial}{\partial x} + E \frac{\partial}{\partial y} + F$$

in the case of the linear equation 9.2.6(1).

The solutions of the corresponding non-homogeneous equation  $Lu = f$  with a given function  $f$  on the right hand side form an affine space. Indeed, if  $Lu_1 = f$ ,  $Lu_2 = f$ ,  $Lu_3 = 0$ , then clearly  $L(u_1 - u_2) = 0$  while  $L(u_1 + u_3) = f$ . Thus, if we succeed to find a single solution to  $Lu = f$ , then

we can add any general solution to the homogeneous equation to obtain a general solution.

Let us illustrate our observation on some of our basic examples. The non-homogenous wave equation  $u_{xx} - u_{yy} = x + y$  has got the general solution

$$u(x, y) = \frac{1}{6}(x^3 - y^3) + f(x - y) + g(x + y)$$

depending on two twice differentiable functions.

The non-homogeneous Laplace equation is called the *Poisson equation*. A general complex valued solution of the Poisson equation  $u_{xx} + u_{yy} = x + y$  is

$$u(x, y) = \frac{1}{6}(x^3 + y^3) + f(x - iy) + g(x + iy)$$

depending on two holomorphic functions  $f$  and  $g$ .

**9.2.8. Separation of variables.** As we have experienced, a straightforward attempt to get solutions is to expect them in a particular simple form. The method of *separation of variables* is based on the assumption that the solution will appear as a product of single variable functions in all variables in question. Let us apply this method on our three special examples.



**Diffusion equation.** We expect to find a general solution of  $\kappa u_{xx} = u_t$  in the form  $u(x, t) = X(x)T(t)$ . Thus the equation says  $\kappa X''(x)T(t) = T'(t)X(x)$ . Assume further  $u \neq 0$  and divide this equation by  $u = XT$ :

$$\frac{X''(x)}{X(x)} = \frac{T'(t)}{\kappa T(t)}.$$

Now the crucial observation comes. Notice the terms on the left and right are function of different variables and thus the equation may be satisfied only if both the sides become constant. We shall have to distinguish the signs of this *separation constant*, so let us write it as  $-\alpha^2$  (choosing the negative option). Thus we have to solve two independent ODEs

$$X'' + \alpha^2 X = 0, \quad T' + \alpha^2 \kappa T = 0.$$

The general solutions are

$$X(x) = A \cos \alpha x + B \sin \alpha x$$

$$T(t) = C e^{-\alpha^2 \kappa t}$$

with free real constants  $A, B, C$ . When combining these solutions in the product, we may absorb the constant  $C$  into the other ones and thus we arrive at the general solution

$$u(x, t) = (A \cos \alpha x + B \sin \alpha x) e^{-\alpha^2 \kappa t}.$$

This solution depends on three real constants.

If we choose a positive separation constant instead, i.e.  $\lambda^2$ , there will be a sign change in our equations and the resulting general solution is

$$u(x, t) = (A \cosh \alpha x + B \sinh \alpha x) e^{\alpha^2 \kappa t}.$$

If the separation constant vanishes, then we obtain just  $u(x, t) = A + Bx$ , independent of  $t$ .



**The Laplace equation.** Assume  $u(x, y) = X(x)Y(y)$  satisfies the equation  $u_{xx} + u_{yy} = 0$  and proceed exactly as above. Thus,  $X''Y + Y''X = 0$  and dividing by  $XY$  and choosing the separation constant  $\alpha^2$ , we arrive at

$$X'' = \alpha^2 X, \quad Y'' = -\alpha^2 Y.$$

The general solution depends on four real constants  $A, B, C, D$

$$u(x, y) = (A \cosh \alpha x + B \sinh \alpha x)(C \cos \alpha y + D \sin \alpha y).$$

If the separation constant is negative, i.e.  $-\alpha^2$ , the roles of  $x$  and  $y$  swap.

**The wave equation.** Let us look how the method works if there are more variables there. Consider a solution  $u(x, y, z, t) = X(x)Y(y)Z(z)T(t)$  of the 3D wave equation

$$\frac{1}{c^2} u_{tt} = u_{xx} + u_{yy} + u_{zz}.$$

Playing the same game again, we arrive at the equation  $\frac{1}{c^2} T'' XYZ = X'' YZT + Y'' XZT + Z'' XYT$ . Dividing by  $u \neq 0$ ,

$$\frac{1}{c^2} \frac{T''}{T} = \frac{X''}{X} + \frac{Y''}{Y} + \frac{Z''}{Z}$$

and since all the individual terms depend on different single variables, they have to be constant. Again, we shall have to keep attention to the signs of the separation constants. For instance, let us choose all constants negative and look at the individual four ODEs

$$\frac{1}{c^2} \frac{T''}{T} = -\alpha^2, \quad \frac{X''}{X} = -\beta^2, \quad \frac{Y''}{Y} = -\gamma^2, \quad \frac{Z''}{Z} = -\delta^2$$

with the constants satisfying  $-\alpha^2 = -\beta^2 - \gamma^2 - \delta^2$ . The general solution is  $u(x, y, z, t) = X(x)Y(y)Z(z)T(t)$  with linear combinations

$$\begin{aligned} T(t) &= A \cos \alpha t + B \sin \alpha t \\ X(x) &= C \cos \beta x + D \sin \beta x \\ Y(y) &= E \cos \gamma y + F \sin \gamma y \\ Z(z) &= G \cos \delta z + H \sin \delta z \end{aligned}$$

with eight real constants  $A$  through  $H$ .

If we choose any of the separation constants positive, the corresponding component in the product would display hyperbolic sine and cosine instead. Of course, the relation between the constants sees the signs as well.

We can also work with complex valued solutions and choose the exponentials as our building blocks (i.e.  $X(x) = e^{\pm i\beta x}$  or  $X(x) = e^{\pm\beta x}$ , etc). For instance, take one of the solutions with all the separation constants negative

$$u(x, y, z, t) = e^{i\beta x} e^{i\gamma y} e^{i\delta z} e^{-i\alpha t} = e^{i(\beta x + \gamma y + \delta z - \alpha t)}.$$

Similarly to the 1D situation, we can again see a "plane wave" propagating along the direction  $(\beta, \gamma, \delta)$  with angular frequency  $\alpha$ .



**9.2.9. Boundary conditions.** We continue with our examples of second order equations and discuss the three most common boundary conditions for them. Let us consider a domain  $\Omega \subset \mathbb{R}^n$ , bounded or unbounded, and a differential operator  $L$  defined on (real or complex valued) functions on  $\Omega$ . We write  $\partial\Omega$  for the boundary of  $\Omega$  and assume this is a smooth manifold.

Locally, such a submanifold in  $\mathbb{R}^n$  is given by one implicit function  $\rho : \mathbb{R}^n \rightarrow \mathbb{R}$  and the unit normal vector  $\nu(x)$ ,  $x \in \partial\Omega$ , to the hypersurface  $\partial\Omega$  is given by the normalized gradient

$$\nu(x) = \frac{\nabla\rho(x)}{\|\nabla\rho(x)\|}.$$

We say that a function  $u$  is differentiable on  $\Omega$ , if it is differentiable on its interior and the directional derivatives  $D_\nu^1 u(x)$  exist in all points of the boundary. Typically we write  $\frac{\partial}{\partial\nu}$  for the derivative in the normal direction.

For simplicity, let us restrict ourselves to  $L$  of the form

$$L = A(x, y) \frac{\partial^2}{\partial x^2} + 2B(x, y) \frac{\partial}{\partial x \partial y} + C(x, y) \frac{\partial^2}{\partial y^2}$$

and look at the equation  $Lu = F(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$ .

#### CAUCHY BOUNDARY PROBLEM

At each point of the boundary  $x \in \partial\Omega$  we prescribe both the value  $\varphi(x) = u(x)$  and the derivative  $\psi(x) = \frac{\partial u}{\partial\nu}(x)$  in the normal unit direction.

The *Cauchy problem* is to solve the equation  $Lu = F$  on  $\Omega$ , subject to  $u = \varphi$  and  $\frac{\partial u}{\partial\nu} = \psi$  on  $\partial\Omega$ .

We shall see that the Cauchy problems very often lead locally to unique solutions, subject to certain geometric conditions on the boundary  $\partial\Omega$ . At the same time, it is often not the convenient setup for practical problems. We shall illustrate this phenomenon on the 2D Laplace equation in the next but one paragraph.

An even simpler possibility is to request only the condition on the values of  $u$  on the boundary  $\partial\Omega$ . Another possibility, often needed in direct applications, is to prescribe the derivatives only. We shall see, that this is reasonable for the Laplace and Poisson equations.

#### DIRICHLET AND NEUMANN BOUNDARY PROBLEMS

At each point of the boundary  $x \in \partial\Omega$  we prescribe the value  $\varphi(x) = u(x)$  or the derivative  $\psi(x) = \frac{\partial u}{\partial\nu}(x)$  in the normal unit direction.

The *Dirichlet problem* is to solve the equation  $Lu = F$  on  $\Omega$ , subject to the condition  $u = \varphi$  on  $\partial\Omega$ .

The *Neumann problem* is to solve the equation  $Lu = F$  on  $\Omega$ , subject to the condition  $\frac{\partial u}{\partial\nu} = \psi$  on  $\partial\Omega$ .

**9.2.10. Uniqueness for Poisson equations.** Because the proof of the next theorem works in all dimensions  $n \geq 2$ , we

shall formulate it for the general Poisson equation

$$(1) \quad \Delta u = \left( \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2} \right) u = F(x_1, \dots, x_n).$$

**Theorem.** Assume  $u$  is a twice differentiable solution of the Poisson equation (1) on a domain  $\Omega \subset \mathbb{R}^n$ . If  $u$  satisfies the Dirichlet condition  $u = \varphi$  on  $\partial\Omega$ , then  $u$  is the only solution of the Dirichlet problem.

If  $u$  satisfies the Neumann condition  $\frac{\partial u}{\partial \nu} = \psi$  on  $\partial\Omega$ , then  $u$  is the unique solution of the Neumann problem, up to an additive constant.

The proof of this theorem relies on a straightforward consequence of the divergence theorem. Remind 9.1.14, saying that for each vector field  $X$  on a domain  $\Omega \subset \mathbb{R}^n$  with hypersurface boundary  $\partial\Omega$

$$(2) \quad \int_M \operatorname{div} X \, dx_1 \dots dx_n = \int_{\partial\Omega} X \cdot \nu \, d\partial\Omega,$$

where  $\nu$  is the oriented (outward) unit normal to  $\partial\Omega$  and  $d\partial\Omega$  stays for the volume inherited from  $\mathbb{R}^n$  on  $\partial\Omega$ .

#### 1ST AND 2ND GREEN'S IDENTITY

**Lemma.** Let  $M \subset \mathbb{R}^n$  be a  $n$ -dimensional manifold with boundary hypersurface  $S$ , and consider two differentiable functions  $\varphi$  and  $\psi$ . Then

$$(3) \quad \int_M (\varphi \Delta \psi + \nabla \varphi \cdot \nabla \psi) \, dx_1 \dots dx_n = \int_S \varphi \nabla \psi \cdot \nu \, dS.$$

This version of the divergence theorem is called the 1st Green's identity.

Next, let us consider one more differentiable function  $\mu$  and  $X = \varphi \mu \nabla \psi - \psi \mu \nabla \varphi$ . The divergence theorem yields the so called 2nd Green's identity

$$(4) \quad \int_M \varphi (\nabla \cdot (\mu \nabla)) \psi - \psi (\nabla \cdot (\mu \nabla)) \varphi \, dx_1 \dots dx_n = \int_S \mu (\varphi \nabla \psi - \psi \nabla \varphi) \cdot \nu \, dS,$$

where  $\nabla \cdot (\mu \nabla)$  means the formal scalar product of the two vector valued differential operators.

**PROOF OF THE GREEN'S IDENTITIES.** The first claim follows by applying (2) to  $X = \varphi \nabla \psi$ , where  $\varphi$  and  $\psi$  are differentiable functions and  $\nabla \psi$  is the gradient of  $\psi$ . Indeed,

$$\begin{aligned} i_{X\omega_{\mathbb{R}^n}} &= \varphi (\nabla \psi \cdot \nu) dS \\ \operatorname{div} X &= \varphi \Delta \psi + \nabla \varphi \cdot \nabla \psi, \end{aligned}$$

where the dot in the second term denotes the scalar product of the two gradients. Let us also notice that the scalar product  $\nabla \psi \cdot \nu$  is just the derivative of  $\psi$  in the direction of the oriented unit normal  $\nu$ .

The second identity is computed the same way and the two terms with the scalar products of two gradients cancel each other. The reader should check the details.  $\square$

**Remark.** A special case of the 2nd Green's identity is worth mentioning. Namely, if  $\mu = 1$  and both  $\psi$  and  $\varphi$  vanish on the boundary  $\partial\Omega$ , we obtain

$$\int_{\Omega} \varphi \Delta \psi - \psi \Delta \varphi \, dx_1 \dots dx_n = 0.$$

This means that the Laplace operator is self-adjoint with respect to the  $L_2$  scalar product on such functions.

**PROOF OF THE UNIQUENESS.** Assume  $u_1$  and  $u_2$  are solutions of the Poisson equation on  $\Omega$ , thus  $u = u_1 - u_2$  is a solution of the homogeneous Laplace equation,

$$\Delta u = \Delta u_1 - \Delta u_2 = F - F = 0.$$

At same time, either  $u = u_1 - u_2 = 0$  on  $\partial\Omega$  or  $\frac{\partial u}{\partial \nu} = 0$  on  $\partial\Omega$ .

Now we exploit the first Green's identity (3) with  $\varphi = \psi = u$ ,

$$\int_{\Omega} (u \Delta u + \nabla u \cdot \nabla u) \, dx_1 \dots dx_n = \int_{\partial\Omega} u \frac{\partial u}{\partial n} \, dS.$$

In both problems, Dirichlet or Neumann, the right hand side vanishes. The first term in the left hand integrand vanishes, too. We conclude

$$\int_{\Omega} \|\nabla u\|^2 \, dx_1 \dots dx_n = 0,$$

but this is possible only if  $\nabla u = 0$  since the integrand is continuous. Thus,  $u = u_1 - u_2$  is constant. But if we solve a Dirichlet problem, then  $u_1$  and  $u_2$  coincide on the boundary and thus they are equal.  $\square$

**9.2.11. Well posed problems.** Consider the Cauchy boundary problem for  $u_{xx} + u_{yy} = 0$ ,  $\partial\Omega$  given by  $y = 0$  and



$$\varphi(x) = u(x, 0) = A_{\alpha} \sin \alpha x$$

$$\psi(x) = u_y(x, 0) = B_{\alpha} \sin \alpha x$$

with the scalar coefficients  $A_{\alpha}$  and  $B_{\alpha}$  depending on the chosen frequency  $\alpha$ . Simple inspection reveals, that we can find such a solution within the result from the separation method:

$$u(x, y) = (A_{\alpha} \cosh \alpha y + \frac{1}{\alpha} B_{\alpha} \sinh \alpha y) \sin \alpha x.$$

Now, choose  $B_{\alpha} = 0$  and  $A_{\alpha} = \frac{1}{\alpha}$ , i.e.

$$u(x, y) = \frac{1}{\alpha} \cosh \alpha y \sin \alpha x.$$

Obviously, when moving  $\alpha$  towards infinity, the Cauchy boundary conditions can become arbitrarily small and still small change of  $B_{\alpha}$  causes arbitrarily big increase of the values of  $u$  in any close vicinity of the line  $y = 0$ .

Imagine, the equation describes some physical process and the boundary conditions reflect some measurements, including some periodic small errors. The results will be horribly instable with respect to these errors in the derivatives. We should admit that the problem is in some sense ill-posed, even locally. This motivates the following definition.

WELL-POSED AND ILL-POSED BOUNDARY PROBLEMS

The problem  $Lu = F$  on the domain  $\Omega$  with boundary conditions on  $\partial\Omega$  is called *well-posed* if all three conditions hold true:

- (1) The boundary problem has got a solution  $u$  (a classical solution means  $u$  is twice continuously differentiable);
- (2) the solution  $u$  is unique;
- (3) the solution is stable with respect to initial data, i.e. “small” change of the boundary conditions results in a “small” change of the solution.

The problem is called *ill-posed*, if any of the above conditions fails.

Usually, the stability in the third condition means that the solution is continuously dependent on the boundary conditions in a suitable topology on the chosen space of functions.

Also the uniqueness required in the second condition has to be taken reasonably. For instance, only uniqueness up to some additive constant makes sense for the Neumann problems.

**9.2.12. Quasilinear equations.** Now we exploit our experience and focus on the (local) Cauchy type problems for equations of arbitrary order. Similarly to the ODEs, we shall deal with problems, where the highest order derivatives are prescribed (more or less) explicitly and the initial conditions are given on a hypersurface up to the order  $k - 1$ .



Some notation will be useful. We shall use the multi-indices to express multivariate polynomials and derivatives, cf. 8.1.15. Further we shall write  $\nabla^k u = \{\partial_\alpha u; |\alpha| = k\}$  for the vector of all derivatives of order  $k$ . In particular,  $\nabla u$  means again the gradient vector of  $u$ .

Some notation will be useful. We shall use the multi-indices to express multivariate polynomials and derivatives, cf. 8.1.15. Further we shall write  $\nabla^k u = \{\partial_\alpha u; |\alpha| = k\}$  for the vector of all derivatives of order  $k$ . In particular,  $\nabla u$  means again the gradient vector of  $u$ .

QUASI-LINEAR PDES

For unknown scalar function  $u$  on a domain  $\Omega \subset \mathbb{R}^n$  we prescribe its derivatives

$$(1) \quad \sum_{|\alpha|=k} a_\alpha(x, u, \dots, \nabla^{k-1}u) \partial_\alpha u = b(x, u, \dots, \nabla^{k-1}u),$$

where  $b$  and  $a_\alpha$  are functions on the tubular domain  $\Omega \times \mathbb{R}^N$ , accomodating all the derivatives, with at least one of  $a_\alpha$  non-zero. We call such equations the (scalar) *quasilinear partial differential equations* (PDE) of order  $r$ .

We call (1) *semilinear* if all  $a_\alpha$  do not depend on  $u$  and its derivatives (thus all the non-linearity hides in  $b$ ).

The *principal symbol* of a semi-linear PDE of order  $k$  is the symmetric  $k$ -linear form  $P$  on  $\Omega$ ,

$$P(x) : (\mathbb{R}^n)^k \rightarrow \mathbb{R}, \quad P(x, \xi, \dots, \xi) = \sum_{|\alpha|=k} a_\alpha(x) \xi^\alpha.$$

For instance, the Poisson equation  $\Delta u = f(x, y, u, \nabla u)$  on  $\mathbb{R}^2$  is a semi-linear equation and its principal symbol is the positive definite quadratic form  $P(\zeta, \eta) = \zeta^2 + \eta^2$ , independent of  $(x, y)$ .

The diffusion equation  $\frac{\partial u}{\partial t} = \Delta u$  on  $\mathbb{R}^3$  has got the symbol  $P(\tau, \zeta, \eta) = \zeta^2 + \eta^2$ , i.e. a positive semi-definite quadratic form, while the wave equation  $\square u = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0$  has got the indefinite symbol  $P(\tau, \zeta) = \tau^2 - \zeta^2$  on  $\mathbb{R}^2$ .

We shall focus on the scalar equations and reduce the problem to a special situation which allows a further reduction to a system of first order equations (quite similarly to the ODE theory). Thus we extend the previous definition to systems of equations. Notice, these are systems of the first kind mentioned in 9.2.4.

#### SYSTEMS OF QUASI-LINEAR PDES

A system of quasi-linear PDEs determines a vector valued function  $u : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ , subject to the vector equation

$$(2) \quad A(x, u, \dots, \nabla^{k-1}u) \cdot \nabla^k u = b(x, u, \dots, \nabla^{k-1}u).$$

Here  $A$  is a matrix of type  $m \times M$  with functions  $a_{i,\alpha} : \Omega \times \mathbb{R}^N$  as entries,  $M = \binom{n+k-1}{k}$  is the number of  $k$ -combinations with repetition from  $n$  objects,  $\nabla^k u$  is the vector of vectors of all the  $k$ th-order derivatives of the components of  $u$ ,  $b : \Omega \times \mathbb{R}^N \rightarrow \mathbb{R}^m$ , and  $\cdot$  means the scalar products of the individual rows in  $A$  with the vectors  $\nabla^k u_i$  of the individual components of  $u$ , matching the individual components in  $b$ .

**9.2.13. Cauchy data.** Next, we have to clarify the boundary condition data. Let us consider a domain  $U \subset \mathbb{R}^n$  and a smooth hypersurface  $\Gamma \subset U$ , e.g.  $\Gamma$  given by an implicit equation  $f(x_1, \dots, x_n) = 0$  locally. Consider the unit normal vector  $\nu(x)$  at each point  $x \in \Gamma$  (i.e.  $\nu = \frac{1}{\|\nabla f\|} \nabla f$  if given implicitly). We would like to find minimal data along  $\Gamma$  determining a solution of 9.2.12(1), at least locally around a given point.

To make things easy, let us first assume that  $\Gamma$  is prescribed by  $x_n = 0$ . Then  $\nu(x) = (0, \dots, 0, 1)$  at all  $x \in \Gamma$  and knowing the restriction of  $u$  to  $\Gamma$ , we also now all derivatives  $\partial_\alpha$  with  $\alpha = (\alpha_1, \dots, \alpha_{n-1}, 0)$ ,  $0 \leq |\alpha|$ . Thus, we have to choose reasonably differentiable functions  $c_j$  on  $\Gamma$ ,  $j = 0, \dots, k-1$ , and posit for all  $j$

$$\partial_\alpha u(x) = c_j(x), \quad \alpha = (0, \dots, 0, j), \quad x \in \Gamma.$$

All the other derivatives  $\partial_\alpha u$  on  $\Gamma$ ,  $0 \leq |\alpha| < \infty$  with  $\alpha_n < k$  are computed inductively by the symmetry of partial derivatives.

Moreover, if  $a_{(0, \dots, 0, k)} \neq 0$ , we can establish the remaining  $k$ -th order derivative by means of the equation 9.2.12(1) and hope to be able to continue inductively. Indeed, writing  $a = a_{(0, \dots, 0, k)}(x, u, \dots, \nabla^{k-1}(u)) \neq 0$  (and similarly leaving out the arguments of the other functions  $a_\alpha$ ), the equation 9.2.12(1) can be rewritten as

$$(1) \quad \frac{\partial^k}{\partial x_n^k} u = \frac{1}{a} \left( - \sum_{|\alpha|=k, \alpha_n \neq k} a_\alpha \partial_\alpha u + b(x, u, \dots, \nabla^{k-1}u) \right).$$

Now, on  $\Gamma$  we can use the already known derivatives to compute directly all the  $\partial_\alpha u$  with  $\alpha_n < k+1$ . But differentiating



the latter equation by  $\frac{\partial}{\partial x_n}$  we obtain the missing derivative of order  $k + 1$  from the known quantities on the right-hand side. By induction, we obtain all the derivatives, as requested.

In the general situation we can iterate the derivative  $D_{\nu(x)}^1 u$  of  $u$  in the direction of the unit normal vector  $\nu$  to the hypersurface  $\Gamma$ :

#### CAUCHY DATA FOR SCALAR PDE

The (smooth or analytic) *Cauchy data* for the  $k$ th order quasi-linear PDE 9.2.12(1) consist of a hypersurface  $\Gamma \subset U$  and  $k$  (smooth or analytic) functions  $c_j$ ,  $0 \leq j \leq k - 1$ , prescribing the derivatives in the normal directions to  $\Gamma$

$$(2) \quad (D_{\nu(x)}^1)^j u(x) = c_j(x), \quad x \in \Gamma.$$

A normal direction  $\nu(x)$ ,  $x \in \Gamma$ , is called *characteristic* for the given Cauchy data, if

$$(3) \quad \sum_{|\alpha|=k} a_\alpha(x, u, \dots, \nabla^{k-1} u) \nu(x)^\alpha = 0.$$

The Cauchy data are called *non-characteristic* if there are no characteristic normals to  $\Gamma$ .

Notice the situation simplifies for the semi-linear equations. Then the characteristic directions do not depend on the chosen functions  $c_j$  from the Cauchy data and they are directly related to the properties of the principal symbol of the equation. In the case of the hyperplane  $\Gamma = \{x_n = 0\}$  treated above, the Cauchy data are non-characteristic if and only if  $a_{(0, \dots, 0, k)} \neq 0$ .

For instance, semi-linear equations of first order always admit characteristic directions since their principal symbols are linear forms and so they must have non-trivial kernels (hyperplanes of characteristic directions). In the three second order examples of the Laplace equation, diffusion equation, and wave equation very different phenomena occur. Since the symbol of the Laplace equation is a positive definite quadratic form, characteristic directions can never appear, independently of our choice of  $\Gamma$ . On the contrary, there are always non-trivial characteristic directions in the other two cases.

#### CHARACTERISTIC CONES OF SEMI-LINEAR PDES

The characteristic directions of a semi-linear PDE on a domain  $\Omega \subset \mathbb{R}^n$  generate the *characteristic cone*  $\mathcal{C}(x) \subset T_x \Omega$  in the tangent bundle,

$$\mathcal{C}(x) = \{\xi \in T_x \Omega; P(x)(\xi, \dots, \xi) = 0\}.$$

The Cauchy data on a hypersurface  $\Gamma$  are non-characteristic if and only if  $(T\Gamma)^\perp \cap \mathcal{C} = \{0\}$ , i.e. the orthogonal complements to the tangent spaces to  $\Gamma$  with respect to the standard scalar product on  $\mathbb{R}^n$  never meet the characteristic cone.

Notice, cones for linear forms are hyperplanes in the tangent space, quadratic cones appear with second order, etc. The tangent vectors to characteristics of the first order quasi-linear equations (as introduced in 9.2.2) are orthogonal to the characteristic normals. We have learned that the first order equations propagate the solutions along the characteristic

lines and so we are not free to prescribe the Cauchy data for the solution in such a case.

**9.2.14. Cauchy-Kovalevskaya Theorem.** As seen so many times already, the analytic mappings are very rigid and most questions related to them boil down to some estimates and smart combinatorial ideas. It is time to remind what happens for analytic equations and Cauchy data in the very special case of the ODEs.



For a single scalar autonomous ODE of first order, the Cauchy data consist of a single point “hypersurface”  $\Gamma = \{x\}$  in  $\Omega \subset \mathbb{R}$  and the value  $u(x)$ . In particular, the Cauchy data are always non-characteristic in dimension one. Already in 6.2.22 we gave a complete proof that the induced derivatives of  $u$  provide a converging power series and thus the only solution, on certain neighborhood of  $x$ . In 8.3.13 we extended the same proof to autonomous systems of ODEs, which verified the same phenomenon for general systems of ODEs of any order  $k$ . Here the Cauchy data again consist of the only point in  $\Gamma$  and all derivatives of  $u$  of orders less than  $k$  (and again, they are always non-characteristic).

In subsequent paragraphs we shall comment on how to extend the ODE proof to the following very famous theorem. In particular, the statement says that we have to expect general solutions to  $k$ th order scalar equations in  $n$  variables to depend on  $k$  independent functions of  $n - 1$  variables. This is in accordance with our experience from simple examples.

CAUCHY-KOVALEVSKAYA THEOREM

**Theorem.** *The analytic Cauchy problem consisting of quasi-linear equation 9.2.12(1) with analytic coefficients and right hand side, and analytic non-characteristic Cauchy data 9.2.13(2) has got a unique analytic solution on a neighborhood of each point in  $\Gamma$ .*

Notice that we have computed explicitly the formal power series for the solution (by an inductive procedure) for the special case when  $\Gamma$  is defined by  $x_n = 0$ . In this case, the theorem claims that this formal series always converges with non-trivial radius of convergence.

The full proof is very technical and we do not have space to bother the readers with all details. In the next paragraphs, we shall provide indications toward the steps in the proof. If the track (or interest) will be lost, the reader should rather jump to 9.2.18.

**9.2.15. Flattening the Cauchy data.** The first step in the proof is to transform the non-characteristic data to the “flat” hypersurface  $\Gamma$  discussed in the beginning of 9.2.13. Remind that for such  $\Gamma$  the non-characteristic condition in 9.2.13(3) reads  $a_{(0,\dots,0,k)} \neq 0$ .



Let us start with the general equation and its analytic Cauchy data on an analytic  $\Gamma \subset \mathbb{R}^n$  (we omit the arguments



of all the functions and  $\ell = 0, \dots, k-1$ )

$$(1) \quad \sum_{|\alpha|=k} a_\alpha \partial_\alpha u = b, \quad \frac{\partial^\ell u}{\partial n^\ell}(x) = c_\ell(x), \quad x \in \Gamma.$$

We shall work locally around some unspecified fixed point in  $\Gamma$ . Since  $\Gamma$  is an analytic hypersurface in  $\mathbb{R}^n$ , there are new local coordinates  $y = \Psi(x)$ , such that

$$\Gamma = \{x; \Psi_n(x) = 0\}.$$

Moreover,  $\Psi$  can be chosen again analytic. Thus, the unit normal vector  $\nu$  to  $\Gamma$  equals to  $\nabla \Psi_n$ , up to a multiple  $\mu^{-1}$  at each point of  $\Gamma$ .

Let  $\Phi = \Psi^{-1}$ , i.e.  $x = \Phi(y)$ , and  $v(y) = u(\Phi(y))$ . Then  $\Phi$  is analytic and the equation transforms to another equation in the coordinates  $y$ ,

$$(2) \quad \sum_{|\alpha|=k} \tilde{a}_\alpha \partial_\alpha v = \tilde{b}$$

with analytic coefficients (they can be all expressed by means of the chain rule and the mutually inverse transformations  $\Phi, \Psi$ ). By the very definition,  $\frac{\partial \Phi}{\partial y_n}$  is a vector (the last column in the matrix  $D^1 \Phi$ ) perpendicular to  $\Gamma$  and thus it must be  $\mu \nu$  (remind the product of the Jacobi matrices  $D^1(\Psi)D^1(\Phi)$  is the identity matrix and the rows  $\nabla \Psi_j$ ,  $j = 1, \dots, n-1$  generate the tangent space  $T\Gamma$ ).

**Claim 1.** The transformed Cauchy data for the equation (2) are analytic.

The hypersurface  $\tilde{\Gamma}$  given by  $y_n = 0$  as well as the coefficients of the equation are analytic. Compute  $\frac{\partial^\ell v}{\partial y_n^\ell}$  on  $\tilde{\Gamma}$ .

$$\begin{aligned} v &= c_0 \circ \Phi \\ \frac{\partial v}{\partial y_n} &= \nabla u \cdot \frac{\partial \Phi}{\partial y_n} = \mu \nabla u \cdot \nu = \mu \frac{\partial u}{\partial \nu} = \mu c_1 \\ \frac{\partial^2 v}{\partial y_n^2} &= \frac{\partial \mu}{\partial y_n} \frac{\partial u}{\partial \nu} + \mu^2 \frac{\partial^2 u}{\partial \nu^2} = \frac{\partial \mu}{\partial y_n} c_1 + \mu^2 c_2 \\ \frac{\partial^3 v}{\partial y_n^3} &= \frac{\partial^2 \mu}{\partial y_n^2} \frac{\partial u}{\partial \nu} + 3\mu \frac{\partial \mu}{\partial y_n} \frac{\partial^2 u}{\partial \nu^2} + \mu^3 \frac{\partial^3 u}{\partial \nu^3} \\ &= \frac{\partial^2 \mu}{\partial y_n^2} c_1 + 3\mu \frac{\partial \mu}{\partial y_n} c_2 + \mu^3 c_3. \end{aligned}$$

Inductively, we see that the transformed functions  $\tilde{c}_j$  are obtained in an analytic way from the functions  $c_i$ ,  $i = 0, \dots, j$ .

**Claim 2.** The Cauchy data for the equation (1) are non-characteristic if and only if the transformed Cauchy data for (2) are non-characteristic, i.e.  $\tilde{a}_{(0, \dots, 0, k)} \neq 0$ .

Compute using the chain rule (remind  $\nabla \Psi_n$  is a vector, the gradient of the last coordinate function of  $\Psi$ , and it is equal to  $\nu$ , up to the non-zero multiple  $\mu^{-1}$ )

$$\begin{aligned} \partial_\alpha u &= \frac{\partial^k v}{\partial y_n^k} (\nabla \Psi_n)^\alpha + \text{lot of terms of lower order in } \frac{\partial}{\partial y_n} \\ &= \mu^{-k} \frac{\partial^k v}{\partial y_n^k} \nu^\alpha + \dots \end{aligned}$$

Substitute into (1)

$$\sum_{|\alpha|=k} a_\alpha \partial_\alpha u = \mu^{-k} \sum_{|\alpha|} a_\alpha \nu^\alpha \frac{\partial^k v}{\partial y_n^k} + \dots$$

We have computed

$$\tilde{a}_{(0,\dots,0,k)} = \mu^{-k} \sum_{|\alpha|=k} a_\alpha \nu^\alpha$$

which is non-zero if and only if the original Cauchy data were non-characteristic.

Since we already verified that all the partial derivatives of  $v$  along  $\tilde{\Gamma}$  can be computed for the non-characteristic Cauchy data with the flat hypersurface  $\tilde{\Gamma}$ , we have actually proved the following claim.

**Proposition.** *The Cauchy data for (1) allow to compute all partial derivatives of its solution  $u$  along the hypersurface  $\Gamma$  if and only if the data are non-characteristic.*

**9.2.16. Reduction to a first-order system.** Without loss of generality, we may consider only the Cauchy data of the form discussed in 9.2.13, i.e. the quasi-linear equation on a domain in  $\mathbb{R}^n$  is

$$(1) \quad \frac{\partial^k}{\partial x_n^k} u = \sum_{|\alpha|=k, \alpha_n \neq k} a_\alpha \partial_\alpha u + b$$

and  $\Gamma$  is given by  $x_n = 0$  with prescribed normal derivatives  $c_0, \dots, c_{k-1}$ . Moreover, we can subtract suitable fixed functions from  $u$  in order to transform the equation into a new one of the same shape and with all the Cauchy data  $c_j$  vanishing on  $\Gamma$ . Indeed, start with  $v = u - c_0$ . This transforms the equation and Cauchy data so that the new  $\tilde{c}_0 = 0$ . If we killed the functions  $\tilde{c}_0, \dots, \tilde{c}_{\ell-1}$ , then we may subtract  $g(x_1, \dots, x_n) = \frac{1}{\ell!} x_n^\ell \tilde{c}_\ell(x_1, \dots, x_{n-1})$  which kills the next one.

The final reduction step is to introduce new functions for all components in the vector  $(u, \nabla u, \dots, \nabla^{k-1} u)$ . Write  $v_1, \dots, v_N$  for all these functions, and add one more function  $v_0(x) = x_n$ . Then we can rewrite our equation (1) as a system of quasi-linear equations of first order on the vector function  $v = (v_0, \dots, v_N)$ .

$$(2) \quad \frac{\partial}{\partial x_n} v_s = \sum_{0 \leq r \leq N} \sum_{i=1}^{n-1} a_{sri} \frac{\partial}{\partial x_i} v_r + b_s, \quad s = 0, \dots, N,$$

where all the coefficients  $a_{sri}$ ,  $b$ , are functions in  $x_1, \dots, x_{n-1}, v_0, \dots, v_N$  and the boundary condition on  $\Gamma = \{x_n = 0\}$  is  $v|_\Gamma = 0$ .

Notice two important facts, the coefficients do not depend on  $x_n$  and all derivatives on the left hand side are  $\frac{\partial}{\partial x_n}$ . This is a technicality which makes the problem similar to the autonomous systems of ODEs.

The principle is obvious from a simple example. Consider 2nd order equation with coefficients  $a_{xx}, a_{xt}, b$

$$u_{tt} = a_{xt} u_{xt} + a_{xx} u_{xx} + b$$

and view  $t, u, u_t, u_x$  as unknown function. Then our equation rewrites as the system of four equations

$$\begin{aligned}\frac{\partial}{\partial t}t &= 1, & \frac{\partial}{\partial t}u &= u_t, & \frac{\partial}{\partial t}u_x &= \frac{\partial}{\partial x}u_t, \\ \frac{\partial}{\partial t}u_t &= a_{xt}\frac{\partial}{\partial x}u_t + a_{xx}\frac{\partial}{\partial x}u_x + b\end{aligned}$$

with boundary condition  $t = u = u_t = u_x = 0$  on the line  $t = 0$ .

**9.2.17. The majorant.** The rest of the proof is very technical, but it is based on the straightforward idea of a majorant for the Cauchy problem 9.2.16(2). Remind we used this method in 8.3.13 when proving the ODE version of the Cauchy-Kovalevskaya theorem.

We shall not go into much detail and only indicate how to generalize the method from 8.3.13. The first step is easy – we have seen already that all derivatives of the solution vector  $v$  at a fixed point  $x$  in  $I$  are computed by chain rule from the Cauchy data. This proves the uniqueness of the analytic solution, if such a solution exists.

Again, it turns out the derivatives are given via universal polynomials in the derivatives of the coefficients  $a_{sri}, b_s$  of the system 9.2.16(2), and the polynomials have got non-negative real coefficients. The reader may easily fill the details as an exercise.

Now, a very similar majorant of the coefficients as in the ODE case can be chosen. First, the analyticity of the coefficients ensures the existence of some suitably small  $r > 0$  and (perhaps big) constant  $C$  such that  $\frac{1}{\alpha!}|\partial_\alpha a_{sri}|r^{|\alpha|} \leq C$ ,  $\frac{1}{\alpha!}|\partial_\alpha b_s|r^{|\alpha|} \leq C$ , and thus

$$|\partial_\alpha a_{sri}| \leq C|\alpha!|r^{-|\alpha|}, \quad |\partial_\alpha b_s| \leq C|\alpha!|r^{-|\alpha|},$$

for all coefficients and multiindices  $\alpha$ .

In particular, all the coefficients can be majorized by the function

$$h(x_1, \dots, x_{n-1}, v_0, \dots, v_N) = \frac{Cr}{r - \sum_{j=1}^{n-1} x_j - \sum_{s=0}^N v_s}.$$

Now, the majorizing system for the vector  $(V_0, \dots, V_N)$  is

$$\frac{\partial V_s}{\partial x_n} = \sum_{0 \leq r \leq N} \sum_{i=1}^{n-1} h \frac{\partial}{\partial x_i} V_r + h, \quad s = 0, \dots, N.$$

Since the coefficients are completely symmetric in the variables, let us expect the solution in the form

$$V_0 = \dots = V_N = W(x_1 + \dots + x_{n-1}, x_n),$$

i.e.  $W$  is a real function of two variables, say  $W(t, y)$ . Substituting into the system we arrive at the linear first order PDE

$$\frac{\partial W}{\partial y} = \frac{Cr}{r - t - NW(t, y)} \left( N(n-1) \frac{\partial W}{\partial t} + 1 \right),$$

with boundary condition  $W(t, 0) = 0$ . The reader may find the solution (e.g. by the method of characteristics)

$$W(t, y) = \frac{1}{Nn} \left( r - t - \sqrt{(r-t)^2 - 2nNCry} \right),$$

a real analytic function on a neighborhood of the origin.

This concludes the proof.<sup>6</sup>

**9.2.18. Back to second order equations.** Let us continue our brief excursion to the PDE world by more detailed comments on the second order semilinear equations. We shall deal with scalar PDEs on a domain  $\Omega \subset \mathbb{R}^n$  with one of the usual boundary conditions.



Thus consider a general linear operator

$$(1) \quad L = \sum_{1 \leq i, j \leq n} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{1 \leq i \leq n} a_i \frac{\partial}{\partial x_i} + a$$

written in coordinates  $(x_1, \dots, x_n)$ . If we consider any other coordinate system  $y = \Phi(x)$ , then the chain rule determines how to transform the equation  $Lu = f$  in coordinates  $x$  into  $\tilde{L}\tilde{v} = \tilde{f}$ , where  $u(x) = v(\Phi(x))$ . Write  $\nabla = (\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n})$  for the gradient operator, dot for the standard scalar product of vectors,  $D^1\Phi$  for the Jacobi matrix of  $\Phi$  in the coordinates  $y$  and  $x$ ,  $\frac{\partial\Phi}{\partial x_i}$  for the  $i$ th column of  $D^1\Phi$ .

$$\begin{aligned} \frac{\partial}{\partial x_i} &= \nabla \cdot \frac{\partial\Phi}{\partial x_i} = \sum_k \frac{\partial\Phi_k}{\partial x_i} \frac{\partial}{\partial y_k} \\ \frac{\partial}{\partial x_i \partial x_\ell} &= \sum_{k, \ell} \frac{\partial\Phi_k}{\partial x_i} \frac{\partial\Phi_\ell}{\partial x_j} \frac{\partial^2}{\partial y_k \partial y_\ell} + \sum_k \frac{\Phi_k}{\partial x_i \partial x_j} \frac{\partial}{\partial y_k}. \end{aligned}$$

Thus, the operator (1) transforms into

$$\tilde{L} = \sum_{1 \leq i, j, k, \ell \leq n} a_{ij} \frac{\partial\Phi_k}{\partial x_i} \frac{\partial\Phi_\ell}{\partial x_j} \frac{\partial^2}{\partial y_k \partial y_\ell} + \sum_{1 \leq i, k \leq n} a_i \frac{\partial\Phi_k}{\partial x_i} \frac{\partial}{\partial y_k} + a.$$

In particular, the principal symbol transforms pointwise as a quadratic form under the linearized transformation  $D^1\Phi$ .

As we know from the linear algebra, the global behavior of real quadratic forms is classified by their signature, i.e. the number of positive and negative entries in the diagonalized matrix, cf. 4.3.2. This is transferred to the following

CLASSIFICATION OF 2ND ORDER QUASI-LINEAR PDES

Consider a second order quasi-linear operator (1) with the principal symbol  $Q$ . The equations  $Lu = f$  and the operator  $L$  are called

- *elliptic* if  $Q$  is either positive or negative definite
- *hyperbolic* if  $Q$  has got the signature  $(n - 1, 1)$  (or equivalently  $(1, n - 1)$ )
- *parabolic* if  $Q$  is positive semidefinite with rank  $n - 1$  and the equation can be rewritten as  $\frac{\partial}{\partial t} u = \tilde{L}(u)$ , where the principal symbol of  $\tilde{L}$  depends on the remaining variables only.

Notice that we actually did not include all possibilities into the above list. We omitted the *ultra-hyperbolic* case, where the rank of  $Q$  is maximal, with the remaining possibilities of signatures. Further, the parabolic equations could appear with the minus sign at  $\frac{\partial}{\partial t}$  (the so called “backwards

<sup>6</sup>The reader may find detailed exposition in many basic books, for instance see the Chapter 1 of the book "Introduction to partial differential equations" by Gerald B. Folland, Princeton, 1995.

parabolic equations”), and the rank of  $Q$  could also drop by more than one. Most of these cases cannot be seen in low dimensions.

If the coefficients  $a_{ij}$ ,  $a_i$  and  $a$  are constant, then the principal symbol  $Q$  is a constant quadratic form, too, and we may choose just linear transformations  $T = D^1\Phi$  instead of general  $\Phi$ . This will always allow us to transform the quadratic form to its canonical form over the entire domain  $\Omega$ .

Many particular examples are discussed explicitly in the other column, see ??, in particular in low dimensions.

Let us also stress that while in dimension  $n = 2$  we may even locally integrate the necessary linearized transformations into a genuine mapping  $\Phi$  and get the canonical forms of the equations even in more general context, see ??, this is mostly not possible in higher dimensions.



Before coming to a few most important examples, let us look at the characteristic directions in the individual cases. If  $L$  is elliptic, then of course there cannot be any characteristic direction. So the (local) Cauchy problem prescribing the analytic value and first normal derivative along any analytic hypersurface  $\Gamma$  will have a locally converging analytic solution around any fixed point. Unfortunately, we have already encountered that this is not a well posed problem even for the two dimensional Laplace equation, cf. 9.2.11.

On the contrary, there will be a  $(n-1)$ -dimensional cone of characteristic direction at each point of a hyperbolic equation, while the parabolic equations will come equipped with a line of characteristic directions.

**9.2.19. The wave equation.** The wave operator in dimension  $n$  is

$$(1) \quad L = \frac{\partial^2}{\partial t^2} - c^2 \Delta,$$

where  $\Delta$  is the Laplace operator  $\Delta = \frac{\partial^2}{\partial x_1^2} + \dots + \frac{\partial^2}{\partial x_n^2}$ ,  $c^2 > 0$  a real constant. The operator  $L$  lives on domains in  $\mathbb{R}^{n+1}$ .

Let us first return to the 2D wave equation  $u_{tt} = c^2 u_{xx}$ . We know the general solution

$$u(x, t) = f(x - ct) + g(y + ct),$$

cf. 9.2.6, which is the superposition of the forward and backward waves  $u_1(x, 0) = f(x)$  and  $u_2(x, 0) = g(x)$ .

This perfectly matches our general expectation from the Cauchy-Kovalevskaya theorem that general solutions to quasilinear second order equations in two variables should depend on two real single variable functions. Moreover, the characteristic directions are  $x \pm ct = 0$  and thus the line  $\Gamma = \{t = 0\}$  is non-characteristic.

Setting the Cauchy boundary data  $u(x, 0) = \varphi(x)$ ,  $\frac{\partial}{\partial t} u(x, 0) = \psi(x)$  and substituting the general solution, we arrive at

$$\varphi(x) = f(x) + g(x), \quad \psi(x) = -cf'(x) + cg'(x),$$

where dash stays for the derivative of the single variable function, as usual. Thus, for any  $s_0$  and  $s$

$$\frac{1}{c} \int_{s_0}^s \psi(x) dx = -f(s) + g(s) + f(s_0) - g(s_0).$$

Add and subtract the equations

$$2g(s) = \varphi(s) + \frac{1}{c} \int_{s_0}^s \psi(x) dx - f(s_0) + g(s_0)$$

$$2f(s) = \varphi(s) - \frac{1}{c} \int_{s_0}^s \psi(x) dx + f(s_0) - g(s_0).$$

Finally, substitute  $s = x - ct$  into  $f(s)$ , and  $s = x + ct$  into  $g(s)$  in order to get the value of  $u(x, t)$  (notice the integrals add nicely, while the constants depending on the choice of  $s_0$  cancel).

$$(2) \quad u(x, y) = \frac{1}{2}(\varphi(x-ct) + \varphi(x+ct)) + \frac{1}{2c} \int_{x-ct}^{x+ct} \psi(y) dy.$$

This solution is often called the *D'Alembert's solution*.

The formula also reveals the continuous dependence of the solution on the boundary conditions. We may conclude that the Cauchy problem seems to be the right boundary value problem for the wave equation (although we have fully proved that it is well-posed only in the dimension two and the analytic category).

In higher dimensions, the situation is much more complicated. One of useful options is to employ the method of separation of variables, but splitting only the time and space variables. Consider the  $n$ -dimensional wave equation and expect the solution in the form  $u(x, t) = F(x)T(t)$  (now  $x \in \mathbb{R}^n$ ,  $t \in \mathbb{R}$ ). Plugging this into (1) and playing the separation method game, we arrive at two equations

$$\Delta F + \alpha F = 0, \quad T'' + \alpha \frac{1}{c^2} T = 0,$$

where  $\alpha$  is the separation constant (usually we consider either  $\alpha^2$  or  $-\alpha^2$  to fix the types of the equations). The first equation is called the *Helmholtz equation* and we shall come back to it below.

**9.2.20. The diffusion equation.** In general dimension  $n$  the *diffusion operator* is

$$(1) \quad L = \frac{\partial}{\partial t} - \kappa \Delta,$$

$\kappa > 0$ , the diffusion equation is considered on domains in  $\mathbb{R}^n \times \mathbb{R}$ .

Again, let us have a look at the simplest 1D diffusion equation  $u_t = \kappa u_{xx}$ . It describes the diffusion process in a one-dimensional object with diffusivity  $\kappa$  (assumed to be constant here) in time. First of all, let us notice that the usual boundary value prescription of the state at time  $t = 0$  is not matching the assumption of the Cauchy-Kovalevskaya theorem. Indeed, taking  $\Gamma = \{t = 0\}$ , the normal direction vector  $\frac{\partial}{\partial t}$  is characteristic.



The intuition related to the expectation on diffusion problems suggests that Dirichlet boundary data should suffice (we just need the initial state and the diffusion then does the rest), or we can combine them with some Neumann data (if we supply heat at some parts of the boundary). Moreover, the process should not be reversible in time, so we should not expect that the solution would extend across the line  $t = 0$ .

Let us look at a classical example considered already by Kovalevskaya. Posit

$$u(0, x) = g(x) = \frac{1}{1 + x^2}$$

on a neighborhood of the origin (perfect analytic boundary data and equation), and expect  $u$  is a solution of  $u_t = u_{xx}$  in the form  $u(t, x) = \sum_{k, \ell \geq 0} c_{k, \ell} \frac{t^k}{k!} \frac{x^\ell}{\ell!}$ .

The equation obviously implies the relations  $c_{k+1, \ell} = c_{k, \ell+2}$  for all  $k, \ell$ . Further, the power series of  $(1 + x^2)^{-1} = \sum_{\ell} (-1)^\ell x^{2\ell}$  is obtained from the geometric power series with argument  $-x$ , and with  $x^2$  substituted for  $x$  in the end.

Thus, for all  $\ell$ ,  $c_{0, 2\ell+1} = 0$ ,  $c_{0, 2\ell} = (-1)^\ell (2\ell)!$ . By the recurrence,  $c_{k, 2\ell} = (-1)^{k+\ell} (2(k + \ell))!$ .

This is a too quick growth for a converging power series. For example, looking at the terms

$$\frac{1}{k!k!} c_{k, k} = \frac{(4k)!}{k!(2k)!},$$

they grow as fast towards the infinity as the expression  $e^{-k} k^k 8^{2k}$ , by the Stirling formula for the factorial (cf. 6.2.17).

We have learned that there cannot be any analytic solution to our Dirichlet boundary problem at all. This example also shows the relevance of all assumptions in the Cauchy-Kovalevskaya theorem.

**9.2.21. Diffusion via Fourier transform.** Fortunately, another straightforward method helps us to solve



the simplest diffusion equation with Dirichlet data. Let us assume  $u(x, t)$  is a solution of  $u_t = \kappa u_{xx}$ ,  $u(x, 0) = \varphi(x)$ . Remind, the Fourier transform (with respect to  $x$ ) transfers the differentiation  $\frac{\partial}{\partial x}$  to algebraic multiplication by  $ix$ , while the other variable  $t$  remains as parameter.

Thus, the Fourier image  $\tilde{u}(\xi, t)$  must obey

$$\tilde{u}_t = -\kappa \xi^2 \tilde{u}, \quad \tilde{u}(\xi, 0) = \tilde{\varphi}.$$

This is a quite simple ODE problem with the general solution

$$\tilde{u}(\xi, t) = C(\xi) e^{-\kappa t \xi^2},$$

while the initial condition implies that the integration constant is just  $C(\xi) = \tilde{\varphi}$ .

Now remember the relation between the Fourier transform and convolution, 7.2.7 at page 660. The image of the convolution is the product of the images, up to the factor  $\sqrt{2\pi}$ . Thus we shall immediately write down the solution  $u$  with

$t > 0$ , once we find the inverse Fourier image of the Gaussian  $f(\xi) = e^{-\kappa t \xi^2}$ . But Fourier images  $\mathcal{F}(f)$  of Gaussians  $f$  are again Gaussians, up to constant, see ??,

$$\mathcal{F}(e^{-ax^2})(\xi) = \frac{1}{\sqrt{2a}} e^{-\frac{\xi^2}{4a}},$$

with any real constant  $a > 0$ . Thus, we can write for  $t > 0$

$$\tilde{u}(\xi, t) = \mathcal{F}(\varphi)\mathcal{F}\left(\frac{1}{\sqrt{2\kappa t}} e^{-\frac{\xi^2}{4\kappa t}}\right).$$

Finally, we obtain  $u$  as the convolution of the initial condition with the so called *heat kernel* function

$$u(x, t) = \frac{1}{2\sqrt{\pi\kappa t}} \int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{4\kappa t}} \varphi(y) dy.$$

Obviously, the solution depends continuously on the boundary condition. We may imagine it models the dynamics of temperature in an infinite homogeneous bar, with some initial distribution of temperature and no losses or gains of energy in time.

Let us also observe the behavior of the solution for  $t$  close to zero. As mentioned in 7.2.9, the Gaussians with variance converging to zero are a good approximation for the so called Dirac delta functions, and indeed the limit of the convolution for  $t \rightarrow 0_+$  is exactly the function  $\varphi$ , as expected.

We shall come back to such convolution based principles a few pages later, after investigating simpler methods.

**9.2.22. Superposition of the solutions.** A general idea to solve boundary value problems is to take a good supply of general solutions and try to take linear combination of even infinite many of them. This means we consider the solution in a form of a series. The type of the series is governed by the available solutions.



Let us illustrate the method on the diffusion equation discussed above. Imagine we want to model the temperature of a homogeneous bar of length  $d$ . Initially, at time  $t = 0$ , the temperature at all points  $x$  is zero. At one of its ends we keep the temperature zero, while the other end will be heated with some constant intensity. Set the bar as the interval  $x \in [0, d] \subset \mathbb{R}$ , and the domain  $\Omega = [0, d] \times [0, \infty)$ . Our boundary problem is

$$(1) \quad u_t = \kappa u_{xx}, \quad u(x, 0) = 0, \quad u(0, t) = 0, \quad \frac{\partial}{\partial x} u(d, t) = \rho,$$

where  $\rho$  is a constant representing the effect of the heating.

The idea is to exploit the general solutions

$$u(x, t) = (A \cos \alpha x + B \sin \alpha x) e^{-\alpha^2 \kappa t}$$

from 9.2.6 with free parameters  $\alpha$ ,  $A$ , and  $B$ . We want to consider a superposition of such solutions with properly chosen parameters and get the solution to our boundary problem in the form combining Fourier series terms with the exponentials. This approach is often called the *Fourier method*.

The condition  $u(0, t) = 0$  suggests to restrict ourselves to  $A = 0$ . Then,  $u_x(x, t) = B\alpha \cos(\alpha x) e^{-\alpha^2 \kappa t}$ . It seems to be difficult now to guess how to combine such solutions, to get something constant in time, as the Neumann part boundary condition requests. But we can help with a small trick.



There are some further obvious solutions to the equation – those with  $u$  depending on the space coordinate only. We may consider

$$v(x, t) = \rho x$$

and seek further our solution in the form  $u(x, t) + v(x)$ . Then  $u$  must again be a solution of the same diffusion equation (1), but the boundary conditions change to  $u(x, 0) = -\rho x$ ,  $u(0, t) = 0$ ,  $\frac{\partial}{\partial x} u(d, t) = 0$ .

Now, we want

$$u_x(x, t) = B\alpha \cos(\alpha x) e^{-\alpha^2 \kappa t} = 0,$$

i.e. we should restrict to frequencies  $\alpha = \frac{1}{2d} n\pi$ , with odd non-negative integers  $n$ . This has settled the second of the boundary condition. The remaining one is  $u(x, 0) = -\rho x$  which sets the condition on the coefficients  $B$  in the superposition

$$\sum_{k \geq 0} B_{2k+1} \sin\left(\frac{(2k+1)\pi x}{2d}\right) = -\rho x$$

on the interval  $x \in [0, d]$ . This is a simple task of finding the Fourier series of the function  $x$ , which we handled in 7.1.10. Combining all this, we get the requested solution  $u(x, t)$  to our problem:

$$\rho x - 8\rho d \frac{1}{\pi^2} \sum_{k \geq 0} \frac{(-1)^k}{(2k+1)^2} \sin\left(\frac{(2k+1)\pi x}{2d}\right) e^{-\kappa \frac{(2k+1)^2 \pi^2 t}{4d^2}}.$$

Even though our supply of general solutions was not big, superposing countably many of them helped us to solve our problem. Notice the behavior at the heated end. If  $t \rightarrow \infty$ , then the all exponential terms in the sum vanish faster than the very first one, the sine terms are bounded, and thus the entire component with the sum vanishes quite fast. Thus, for big  $t$ , the heated end will increase its temperature nearly linearly with the speed  $\rho$ .

**9.2.23. Separation in transformed coordinates.** As we



have seen several times, it is very useful to view a given equation rather as an independent object expressed in some particular coordinates.

The practical problems mostly include some symmetries and then we should like to find some suitable coordinates in order to see the equation in some simple form.

As an example, let us look at the Laplace operator  $\Delta$  in the polar coordinates in the plane, and cylindrical or spherical coordinates in the space. Writing as usual  $x = r \cos \varphi$ ,  $y = r \sin \varphi$  for the polar transformation, the Laplace operator gets the neat form

$$(1) \quad \Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2} + \frac{1}{r} \frac{\partial}{\partial r} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2}.$$

The reader should perform the tedious but straightforward computation. Similarly,

$$(2) \quad \Delta = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2} + \frac{\partial^2}{\partial z^2},$$

$$(3) \quad \Delta = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin^2 \psi} \frac{\partial^2}{\partial \varphi^2} + \frac{1}{r^2 \sin \psi} \frac{\partial}{\partial \psi} \left( \sin \psi \frac{\partial}{\partial \psi} \right)$$

in the cylindrical and spherical coordinates, respectively.

Let us illustrate the use on the following problem. Imagine a twisted circular drum, whose rim suffers a small vertical displacement. We should model the stabilized position of the drumskin.

Intuitively, we should describe the drumskin position by the  $2D$  wave equation, but since we are interested in the state with  $\frac{\partial}{\partial t}$  vanishing, we actually take  $u$  as the vertical displacement in the interior of the unit circle,  $\Omega = \{x^2 + y^2 \leq 1\} \subset \mathbb{R}^2$  and request  $\Delta u = 0$ , subject to the Dirichlet boundary problem prescribing the vertical displacement  $u(x, y) = f(x, y)$  of the rim.

Obviously, we want to consider the problem in the polar coordinates, where the boundary condition gets the neat form  $u(1, \varphi) = g(\varphi)$ . Say  $g(\varphi) = \varepsilon \sin \varphi + \varepsilon^2 \sin 5\varphi$  with some small constant  $\varepsilon \geq 0$ .

We shall apply the separation of variables method to these data. Expecting the solution in the form  $u(r, \varphi) = R(r)\Phi(\varphi)$ , the equation implies (after dividing by  $R\Phi$ )

$$\frac{R''}{R} + \frac{1}{r} \frac{R'}{R} + \frac{1}{r^2} \frac{\Phi''}{\Phi} = 0.$$

Thus, multiplying by  $r^2$  and considering the separation constant  $\alpha^2$ , we arrive at two ODEs

$$\Phi'' + \alpha^2 \Phi = 0, \quad r^2 R'' + rR' - \alpha^2 R = 0.$$

Fortunately, they are both easy to solve. From the first equation (with  $\alpha > 0$ )

$$\Phi(\varphi) = A \cos \alpha\varphi + B \sin \alpha\varphi,$$

while the other equation transform by  $S(t) = R(\exp t)$  into an equation with constant coefficients and its solution yields

$$R(r) = Cr^\alpha + Dr^{-\alpha}$$

with  $\alpha \neq 0$ . If  $\alpha = 0$ , then the solution for  $R$  is  $R(r) = C \ln r + D$  and the previous equation implies  $\Phi(\varphi) = A\varphi + B$ .

In fact, we insist the solution  $u = R\Phi$  be a single valued function in the plane and so we can allow only integer values of  $\alpha$ , including  $\alpha = 0$  when  $\Phi$  becomes a constant (again, any non-zero multiple would lead to multi-valued solutions  $u$ ). Thus the general solution of the Laplace equation coming from the separation of variables method and superposition is

$$(4) \quad u(r, \varphi) = C_0 \ln r + D_0 + \sum_{n=1}^{\infty} (A_n \cos n\varphi + B_n \sin n\varphi)(C_n r^n + D_n r^{-n}).$$

In our problem, we clearly insist in having  $u$  finite at the origin and thus all  $D_n$  and the  $C_0$  have to vanish. Now we can employ the boundary condition

$$D_0 + \sum_{n=1}^{\infty} c_n (A_n \cos n\varphi + B_n \sin n\varphi) = \varepsilon \sin \varphi + \varepsilon \sin 5\varphi.$$

This is a very simple case of Fourier series and we see immediately that all the coefficients have to vanish except of the  $B_1$  and  $B_5$  and the requested solution is

$$u(r, \varphi) = \varepsilon r \sin \varphi + \varepsilon r^5 \sin 5\varphi.$$

The higher the frequency of the twist of the rim, the slower the distortion develops in the center of the drumskin. The method works for every boundary condition  $u(1, \varphi) = g(\varphi)$ , if we are able to find its Fourier series.

**9.2.24. The Helmholtz equation.** In 9.2.19, we looked for solutions of the nD wave equation  $u_{tt} - c^2 \Delta u = 0$  in the form  $u(x, t) = F(x)T(t)$ , where  $x \in \mathbb{R}^n$ ,  $t \in \mathbb{R}$ . This (partial) separation of variables with negative separation constant  $-\alpha^2$  leads to the *Helmholtz equation*

$$\Delta F + \alpha F = 0,$$

together with the easily solved  $T'' + \alpha^2 T = 0$ .

Let us treat the 2D case in the polar coordinates, again using separation of variables. Thus, the Helmholtz equation gets the form

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2} + \alpha^2 F = 0$$

and we seek for  $F(r, \varphi) = R(r)\Phi(\varphi)$ . Writing  $\beta^2$  for the separation constant now, we arrive at the two equations (the second one is multiplied by  $r^2$ , for convenience)

$$\Phi'' + \beta^2 \Phi = 0, \quad r^2 R'' + rR' + (\alpha^2 r^2 - \beta^2)R = 0.$$

The angular component equation has got the obvious solutions  $A \cos \beta\varphi + B \sin \beta\varphi$ , and again we have to restrict  $\beta$  to integers in order to get single-valued solutions. With  $\beta = m$ , the radial equation is the well known Bessel's ODE of order  $m$  (notice our equation gets the form we had in ?? once we substitute  $z = \alpha r$ ), with the general solution

$$R(r) = CJ_m(\alpha r) + DY_m(\alpha r),$$

where  $J_m$  and  $Y_m$  are the special Bessel functions of the first and second kinds.

We have obtained a general solution which is very useful in practical problems, cf. ??.

**9.2.25. Non-homogeneous equations.** Finally, we add a



few comments on the non-homogeneous linear PDEs. Although we provide arguments for the claims, we shall not go into technical details of proofs because of the lack of space. Still, we hope this limited insight will motivate the reader to seek for further sources to learn more.

As always, facing a problem  $Lu = f$ , we have to find a single particular solution to this problem, and we may then add all solutions to the homogeneous problem  $Lu = 0$ . Thus, if we have to match say Dirichlet conditions  $u = g$  on the boundary  $\partial\Omega$  of a domain  $\Omega$ , and we know some solution  $w$ , i.e.  $Lw = f$  (not taking care of the boundary conditions), then we should find a solution  $v$  to the homogenous Dirichlet problem with the boundary condition  $g - w|_{\partial\Omega}$ . Clearly the sum  $u = v + w$  will solve our problem.

In principle, we may always consider superpositions of known solutions as in the Fourier method above. We shall indicate a more conceptual and general approach now briefly.

Let us come back to the 1D diffusion equation and our solution of a homogeneous problem by means of the Fourier transform in 9.2.21. The solution of  $u_t = \kappa u_{xx}$  with  $u(x, 0) = \varphi$  is a convolution of the boundary values  $u(x, 0)$  with the *heat kernel*

$$(1) \quad Q(x, t) = \frac{1}{\sqrt{4\pi\kappa t}} e^{-\frac{x^2}{4\kappa t}}.$$

Now, the crucial observation is that  $u(x, t) = Q(x, t)$  is a solution to  $L(u) = u_t - \kappa u_{xx} = 0$  for all  $x$  and  $t > 0$ , while on neighborhood of the origin it behaves as the Dirac delta function in the variable  $x$ . (The first part is a matter of direct computation, the second one was revealed in 9.2.21 already.)

The latter observation suggests, how to find the particular solutions to a non-homogeneous problem. Consider the integral of the convolution

$$(2) \quad u(x, t) = \int_0^t \left( \int_{-\infty}^{\infty} Q(x-y, t-s) f(y, s) dy \right) ds.$$

The derivative  $u_t$  will have two terms. In the first one we differentiate with respect to the upper limit of the outer integral, while the other one is the derivative inside the integrals. The derivatives with respect to  $x$  are evaluated inside the integrals. Thus, in the evaluation of  $L = \frac{\partial}{\partial t} - \kappa \frac{\partial^2}{\partial x^2}$  the terms inside of the integral cancel each other (remember  $Q$  is a solution for all  $x$ , and  $t > 0$ ) and only the first term of  $u_t$  survives.

It seems obvious that this term is the evaluation of the integrand with  $s = t$ . Although, these values are not properly defined, we may verify this claim in terms of taking limit  $(t-s) \rightarrow 0_+$ . This leads to

$$\lim_{s \rightarrow t_-} \int_{-\infty}^{\infty} Q(x-y, t-s) f(y, s) dy = f(x, s).$$

Thus, (2) is a particular solution and clearly  $u(x, 0) = 0$ .

The solution of the general Dirichlet problem  $L(u) = f$ ,  $u(x, 0) = \varphi$  on  $\Omega = \mathbb{R} \times [0, \infty)$  is

$$(3) \quad u(x, t) = \int_{-\infty}^{\infty} Q(x-y, t) \varphi(y) dy + \int_0^t \left( \int_{-\infty}^{\infty} Q(x-y, t-s) f(y, s) dy \right) ds.$$

Let us summarize the achievements and try to get generalization to general dimensions.

First, we can generalize the heat kernel function  $Q$  writing its nD variant depending on the distance  $r$  from the origin only. Consider the formula with  $x \in \mathbb{R}^n$  as the product of the 1D heat kernels for each of the variables in  $x$ .

$$(4) \quad Q(x, t) = \frac{1}{\sqrt{(4\pi\kappa t)^n}} e^{-\frac{\|x\|^2}{4\kappa t}}.$$

Then taking the  $n$ -dimensional (iterated) convolution of  $Q$  with the boundary condition  $\varphi$  on the hyperplane  $t = 0$  provides the solution candidate

$$(5) \quad u(x, t) = \int_{\mathbb{R}^n} Q(x-y, t) \varphi(y) dy_1 \dots dy_n.$$

Indeed, a straightforward (but tedious) computation reveals that  $Q$  is a solution to  $L(u) = 0$  in all points  $(x, t)$  with  $t > 0$ , and  $Q$  behaves again as the Dirac delta at the origin. In particular (5) is a solution to the Dirichlet problem  $L(u) = 0$ ,  $u(x, 0) = \varphi$  and we can also obtain the non-homogeneous solutions similarly to the 1D case.

**9.2.26. The Green's functions.** The solutions to the (non-homogeneous) diffusion equation constructed in the last paragraph are built on a very simple idea – we find a solution  $G$  to our equation which is defined everywhere except in the origin and blows up in the origin at the speed making it into a Dirac delta function at the origin. A convolution with such kernel  $G$  is then a good candidate for solutions for . Let us try to mimic this approach for the Laplace and Poisson equations now.

Actually, we shall modify the strategy by requesting

ABOUT 1 PAGE TO BE FINISHED – THE SPHERICAL SYMMETRIC SOLUTION TO LAPLACE  $\Rightarrow$  GREEN'S FUNCTION  $\Rightarrow$  SOLUTION TO POISSON SIMILARLY TO THE DIFFUSION.

### 3. Remarks on Variational Calculus

Many practical problems look for minima or maxima of real functions  $\mathcal{J} : \mathcal{S} \rightarrow \mathbb{R}$  defined on some spaces of functions. In particular, many laws of nature can be expressed as certain “minimum principle” concerning some space of mappings.

The basic idea is exactly the same as in the elementary differential calculus: we aim at finding the best linear approximations of  $\mathcal{J}$  at fixed arguments  $u \in \mathcal{S}$ , we recognize the critical points (those with vanishing linearization), and then we perhaps look at the quadratic approximations at the critical points. However, all these steps are far more intricate, need a lot of care, and may provide nasty surprises.

#### 9.3.1. Simple examples first.

If we know the sizes of tangent vectors to curves, we may ask what is the shortest distance between two points. In the plane  $\mathbb{R}^2$ , this means we have got a quadratic form  $g(x) = (g_{ij}(x))$ ,  $1 \leq i, j \leq 2$ , at each  $x \in \mathbb{R}^2$  and we want to integrate (the dots mean derivatives in time  $t$ ,  $u(t) = (u_1(t), u_2(t))$  are differentiable paths)



$$(1) \quad \mathcal{J}(u) = \int_{t_1}^{t_2} \sqrt{g(u(t))(\dot{u}(t))} dt$$

to get the distance between the two given points  $u(t_1) = (u_1(t_1), u_2(t_1)) = A$  and  $u(t_2) = (u_1(t_2), u_2(t_2)) = B$ . If the size of the vectors is just the Euclidean one, and we consider curves  $u(t) = (t, v(t))$ , i.e., graphs of functions of one variable, the length (1) becomes the well known formula

$$(2) \quad \mathcal{J}(u) = \int_{t_1}^{t_2} \sqrt{1 + \dot{v}(t)^2} dt.$$

Quite certainly we all believe that the minimum for fixed boundary values  $v(t_1)$  and  $v(t_2)$  must be a straight line. But so far, we have not formulated the problem itself. What is the space of curves we deal with? If we allowed non-continuous ones, then shorter paths are available! So we should aim at proving that the lines are the minimal curves among the continuous ones. Do we need them to be differentiable? In some sense we do, since the derivative appears in our formula for  $\mathcal{J}$ , but we need to have the integrand defined only almost everywhere. For example, this will be true for all Lipschitz curves.

In general,  $g(u)(\dot{u}) = g_{11}(u)\dot{u}_1^2 + 2g_{12}(u)\dot{u}_1\dot{u}_2 + g_{22}(u)\dot{u}_2^2$ . Such lengths of vectors are automatically inherited from the ambient Euclidean  $\mathbb{R}^3$  on every hypersurface in the space. Thus, finding the minimum of  $\mathcal{J}$  means finding the shortest track in a real terrain (with hills and valleys).

If we choose a positive function  $\alpha$  on  $\mathbb{R}^2$  and consider  $g(x) = \alpha(x)^2 \text{id}_{\mathbb{R}^2}$ , i.e., the Euclidean size of vectors scaled by  $\alpha(x) > 0$  at each point  $x \in \mathbb{R}^2$ , we obtain

$$(3) \quad \mathcal{J}(u) = \int_{t_1}^{t_2} \alpha(t, v(t)) \sqrt{1 + \dot{v}(t)^2} dt.$$

We can imagine the speed  $1/\alpha$  of a moving particle (or light) in the plane depends of the values of  $\alpha$  (the smaller is  $\alpha$ , the

bigger is the speed) and our problem will be to find the shortest path in terms of the time necessary to pass from  $A$  to  $B$ .

As a warm up, consider  $\alpha = 1$  in the entire plane, except the vertical strip  $V = \{(t, y); t \in [a, a + b]\}$  where  $\alpha = N$  and take  $A = (0, 0)$ ,  $B = (a + b, c)$ ,  $a, b, c > 0$ . We can imagine  $V$  is a lake, you have to get from  $A$  to  $B$  by running and swimming, and you are swimming  $N$  times slower than running. If we believe that the straight lines are the minimizers for constant  $\alpha$ , then it is clear that we have to find the optimal point  $P = (a, p)$  on the bank of the lake where we start swimming. The total time  $T(p)$  will then be ( $s$  is our actual speed when running straight)

$$\frac{|AP|}{s} + \frac{|PB|}{s/N} = \frac{1}{s} \left( \sqrt{p^2 + a^2} + N \sqrt{(c - p)^2 + b^2} \right)$$

and we want to find the minimum of  $T(p)$ . The critical point is given by

$$\frac{p}{\sqrt{p^2 + a^2}} = N \frac{c - p}{\sqrt{(c - p)^2 + b^2}} \implies \sin \varphi = N \sin \psi,$$

where  $\varphi$  is the angle between our running track and the normal to the boundary of  $V$ , while  $\psi$  is the angle between our swimming track and the normal to the boundary (draw a picture yourself!). Thus we have recovered the famous Snell law of light diffraction saying that the proportion of the sine values of the angles is equal to the proportion of the speeds. (Of course, to finish the solution of the problem, the reader should find the solution  $p$  of the quartic equation and check that it is a minimum.)

**9.3.2. Variational problems.** We shall restrict our attention to the following class of problems.

GENERAL FIRST ORDER VARIATIONAL PROBLEMS

Consider an open Riemann measurable set  $\Omega \subset \mathbb{R}^n$ , the space  $C^1(\Omega)$  of all differentiable mappings  $u : \Omega \rightarrow \mathbb{R}^m$ , a  $C^2$  function  $F = F(x, y, p) : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{nm}$  and set the functional

$$(1) \quad \mathcal{J}(u) = \int_{\Omega} F(x, u(x), D^1 u(x)) \omega_{\mathbb{R}^n},$$

i.e.,  $\mathcal{J}(u)$  is computed as the ordinary integral of a Riemann integrable function  $f(x) = F(x, u(x), D^1 u(x))$  where  $D^1 u$  is the Jacobi matrix (the differential) of  $u$ . The function  $F$  is called the *Lagrangian* of the *variational problem* and our task is to find the minimum of  $\mathcal{J}$  and the corresponding *minimizer*  $u$  with prescribed boundary values  $u$  on the boundary  $\partial\Omega$  (and perhaps some further conditions restricting  $u$ ).

Mostly we shall restrict ourselves to the case  $n = m = 1$ , like in the previous paragraph, where  $u$  is a real differentiable function defined on an interval  $(t_1, t_2)$  and the function  $F = F(t, y, p) : \mathbb{R}^3 \rightarrow \mathbb{R}$ ,

$$(2) \quad \mathcal{J}(u) = \int_{\Omega} F(t, u(t), \dot{u}(t)) dt.$$



We saw  $F = \sqrt{1 + p^2}$ ,  $F = \alpha(t)\sqrt{1 + p^2}$  in the previous paragraph. If we take  $F = y\sqrt{1 + p^2}$ , the functional  $\mathcal{J}$  computes the area of the rotational surface given by the graph of the function  $u$  (up to a constant multiple). In all cases we may set the boundary values  $u(t_1)$  and  $u(t_2)$ .

Actually, our differentiability assumptions are too strict as we saw already in our last example above, where  $F$  was differentiable except of the boundary of the lake  $V$ . We can easily extend our space of functions to piecewise differentiable  $u$  and request  $F(t, u(t), \dot{u}(t))$  to be piecewise differentiable for all such  $u$ 's (as always, piecewise differentiable means the one-side derivatives exist at all points).

A maybe shocking example is the following functional:

$$(3) \quad \mathcal{J}(u) = \int_0^1 (\dot{u}(t)^2 - 1)^2 dt$$

on piece-wise differentiable functions on  $[0, 1]$  (i.e.  $F$  is the neat polynomial  $(p^2 - 1)^2$ ). Clearly,  $\mathcal{J}(u) \geq 0$  for all  $u$  and if we set  $u(0) = u(1) = 0$ , then any zig-zag piecewise linear function  $u$  with derivatives  $\pm 1$  satisfying the boundary conditions achieves the zero minimum. At the same time, there is no minimum among the differentiable functions  $u$  (find a quick proof of that!), but we can approximate any of the zig-zag minima by smooth ones at any precision.

**9.3.3. More examples.** Let us develop a general method how to find the analogy to the critical points form the elementary calculus here. We shall find the necessary steps dealing with a specific set of problems in this paragraph. Let us work with the Lagrangian generalizing the previous examples:



$$(1) \quad F(t, y, p) = y^r \sqrt{1 + p^2}$$

$r > 0$ , and write  $F_t, F_y, F_p$ , etc., for the corresponding partial derivatives. Consider the variational problem on an interval  $I = (t_1, t_2)$  with fixed boundary conditions  $u(t_1)$  and  $u(t_2)$  and assume  $u \in C^2(I)$ ,  $u(t) > 0$ . Let us consider any differentiable  $v$  on  $I$  with  $v(t_1) = v(t_2) = 0$  (or even better  $v$  with compact support inside of  $I$ ). Then  $u + \delta v$  fulfills the boundary conditions for all small real  $\delta$ 's and consider

$$\mathcal{J}(u + \delta v) = \int_{t_1}^{t_2} F(t, u(t) + \delta v(t), \dot{u}(t) + \delta \dot{v}(t)) dt.$$

Of course, the necessary condition for  $u$  being a critical point must be  $\frac{d}{d\delta}|_0 \mathcal{J}(u + \delta v) = 0$ , i.e., (remind the derivative with respect to a parameter can be swapped with the integration)

$$(2) \quad 0 = \int_{t_1}^{t_2} F_y(t, u(t), \dot{u}(t))v(t) + F_p(t, u(t), \dot{u}(t))\dot{v}(t) dt.$$

Integrating the second term in (2) per partes immediately yields (remember  $v(t_1) = v(t_2) = 0$ )

$$0 = \int_{t_1}^{t_2} (F_y(t, u(t), \dot{u}(t))v(t) - \frac{d}{dt} F_p(t, u(t), \dot{u}(t)))v(t) dt.$$

This condition will be certainly satisfied if the so called *Euler equation* holds true for  $u$  (we prove this is a necessary condition in lemma 9.3.6)

$$(3) \quad \frac{d}{dt} F_p(t, u(t), \dot{u}(t)) = F_y(t, u(t), \dot{u}(t)).$$

An equivalent form of this equation for  $\dot{u}(t) \neq 0$  is (we omit the arguments  $t$  of  $u$  and  $\dot{u}$ )

$$(4) \quad F_t(t, u, \dot{u}) = \frac{d}{dt} (F(t, u, \dot{u}) - \dot{u} F_p(t, u, \dot{u})).$$

In our case of  $F(t, y, p) = y^r (1 + p^2)^{1/2}$ ,  $F_t$  vanishes identically,  $F_p = y^r p (1 + p^2)^{-1/2}$  and thus, if we further assume  $r \neq 0$ ,  $u > 0$ , the term in the bracket has to be a positive constant  $C^r$ :

$$C^r = u^r (1 + \dot{u}^2)^{1/2} - \dot{u} u^r \dot{u} (1 + \dot{u}^2)^{-1/2} = u^r (1 + \dot{u}^2)^{-1/2}.$$

We have arrived at the differential equation

$$(5) \quad u = C(1 + \dot{u}^2)^{1/2r}$$

which we are going to solve.


Consider the transformation  $\dot{u} = \tan \tau$ , i.e.,

$$u = C(1 + (\tan \tau)^2)^{1/2r} = C(\cos \tau)^{-1/r},$$

and so  $du = \frac{C}{r} (\cos \tau)^{-1/r} \tan \tau d\tau$ . Consequently,  $dt = \frac{1}{\dot{u}} du = \frac{C}{r} (\cos \tau)^{-1/r} d\tau$  and by integration we arrive at the very useful parametrization of the solutions by the parameter  $\tau$  (which is actually the slope of the tangent to the solution graph):

$$(6) \quad t = t_0 + \frac{C}{r} \int_0^\tau (\cos s)^{-1/r} ds \quad u = C(\cos \tau)^{-1/r}.$$

Now, we can summarize the result for several interesting values of  $r$ . First, if  $r = 0$  (which we excluded on the way), then the Euler equation (3) reads



$$\ddot{u}(1 + \dot{u}^2)^{-3/2} = 0,$$

which implies  $\ddot{u} = 0$  and thus the potential minimizers should be straight lines as expected. (Notice that we have not proved yet that the Euler equation is indeed a necessary condition, we shall come to that in the next paragraphs.)

For general  $r \neq 0$ , the Euler equation (3) tells (a straightforward computation!)

$$\ddot{u} = r \frac{1 + \dot{u}^2}{u}$$

and thus the sign of the second derivative coincides with the sign of  $r$ . In particular, the potential minimizers are always concave functions (if  $r < 0$ ) or convex (if  $r > 0$ ).

If  $r = -1$ , the parametrization (6) leads to (an easy integration!)

$$(7) \quad t = t_0 - C \sin \tau, \quad u = C \cos \tau,$$

thus for  $\tau \in [-\pi/2, \pi/2]$  our solutions are half-circles with radius  $C$  in the upper halfplane, centred at  $(t_0, 0)$ .

For  $r = -1/2$ , the solution is

$$(8) \quad t = t_0 - \frac{C}{2} (2\tau + \sin 2\tau), \quad \frac{C}{2} (1 + \cos 2\tau)$$

which is a parametric description of a fixed point on a circle with diameter  $C$  rolling along the  $t$  axis, the so called *cycloid* curve. Now,  $\tau \in [-\pi/2, \pi/2]$  provides  $t$  running from  $t_0 + \frac{1}{2}C\pi$  to  $t_0 - \frac{1}{2}C\pi$ , while  $u$  is zero in the point  $t_0 \pm \frac{1}{2}C\pi$  and reaches the highest point at  $t = t_0$ . (Draw pictures!)

Next, look at  $r = 1/2$ . Another quick integration reveals  $t = t_0 + 2C \tan \tau = t_0 + 2C\dot{u}$ , and we can compute  $\dot{u}$  and substitute into (5) to obtain

$$u = C + \frac{1}{4C}(t - t_0)^2.$$

Thus the potential minimizers are parabolas with the axis of symmetry  $t = t_0$ . If we fix  $A = (0, 1)$  and a  $t_0$ , there are two relevant choices  $C = \frac{1}{2}(1 \pm \sqrt{1 - t_0^2})$  whenever  $|t_0| < 1$  (and no options for  $|t_0| > 1$ ). The two parabolas will have two points of intersection,  $A$  and another point  $B$ . Clearly only one of them should be the actual minimizer. Moreover, the reader could try to prove that the parabola  $u = \frac{1}{4}t^2$  touches all of them and has them all on the left (this is the so called envelope of all the family of parabolas). Thus, there will be no potential minimizer joining the point  $A = (1, 0)$  to an arbitrary point on the right of the parabola  $u = \frac{1}{4}t^2$ .

The last case we come to is  $r = 1$ , i.e., the case of the area of the surface of the rotational body drawn by the graph. Here we better use another parametrization of the slope of the tangent, we set  $\dot{u} = \sinh \tau$ . A very similar computation as above then immediately leads to  $t = t_0 + \frac{C}{r} \int_0^\tau \cosh s \, ds$  and we arrive at the result<sup>7</sup>

$$(9) \quad u(t) = C \cosh \frac{t-t_0}{C}.$$

**9.3.4. Critical points of functionals.** Now we shall develop a bit of theory verifying that the steps done in the previous examples really provided necessary conditions for solutions of the variational problems. In order to underline the essential features, we shall first introduce the basic tools in the realm of general normed vector spaces, see 7.3.1. The spaces of piecewise differentiable functions on an interval with the  $L_p$  norms can serve as typical examples. We shall deal with mappings  $\mathcal{F} : \mathcal{S} \rightarrow \mathbb{R}$  called (real) *functionals*.



#### THE FIRST DIFFERENTIAL

Let  $\mathcal{S}$  be a vector space equipped with a norm  $\| \cdot \|$ . A continuous linear mapping  $L : \mathcal{S} \rightarrow \mathbb{R}$  is called a *continuous linear functional*.

A functional  $\mathcal{F} : \mathcal{S} \rightarrow \mathbb{R}$  is said to have the *differential*  $D_u \mathcal{F}$  at a point  $u \in \mathcal{S}$  if there is a continuous linear functional  $L$  such that

$$(1) \quad \lim_{v \rightarrow 0} \frac{\mathcal{F}(u+v) - \mathcal{F}(u) - L(v)}{\|v\|} = 0.$$

<sup>7</sup>Some more details on the set of examples of this paragraph can be found in the article "Elementary Introduction to the Calculus of Variations" by Magnus R. Hestenes, *Mathematics Magazine*, Vol. 23, No. 5 (May - Jun., 1950), pp. 249-267.

In the very special case of the Euclidean  $\mathcal{S} = \mathbb{R}^n$ , we have recovered the standard definition of the differential, cf. 8.1.7 (just notice that all linear functionals are continuous on a finite dimensional vector space). Again, the differential is computed via the directional derivatives.<sup>8</sup> Indeed, if (1) holds true, then for each fixed  $v \in \mathcal{S}$

$$(2) \quad \delta\mathcal{F}(u)(v) = \lim_{t \rightarrow 0} \frac{\mathcal{F}(u + tv) - \mathcal{F}(u)}{t} = \frac{d}{dt}\bigg|_0 \mathcal{F}(u + tv)$$

exists and  $L(v) = \delta\mathcal{F}(u)(v)$ . We call  $\delta\mathcal{F}(u)$  the *variation of the functional  $\mathcal{F}$  at  $u$* .

A point  $u \in \mathcal{S}$  is called a *critical point* if  $\delta\mathcal{F}(u) = 0$ . We say that  $\mathcal{F}$  has got a *local minimum* at  $u$  if there is an open neighborhood  $U$  of  $u$  such that  $\mathcal{F}(w) \geq \mathcal{F}(u)$  for all  $w \in U$ . Similarly, we define local maxima and talk about local extrema.

If  $u$  is an extreme of  $\mathcal{F}$ , then in particular  $t = 0$  must be an extreme of the function  $\mathcal{F}(u + tv)$  of one real variable  $t$ , where  $v$  is arbitrary. Thus the extremes have to be at critical points, if the variations exist.

Next, let us assume the variations exist at all points in a neighborhood of a critical point  $u \in \mathcal{S}$ . Then, again exactly as in the elementary calculus, considering two increments  $v, w \in \mathcal{S}$  we consider the limit

$$(3) \quad \delta^2\mathcal{F}(u)(v, w) = \lim_{t \rightarrow 0} \frac{\delta\mathcal{F}(u + tv)(w) - \delta\mathcal{F}(u)(w)}{t}.$$

If the limits exist for all  $u, v$ , then clearly  $\delta^2\mathcal{F}(u)$  is a bilinear mapping. Then,  $\delta^2\mathcal{F}(u)(w, w)$  is a quadratic form which we can consider as a second order approximation of  $\mathcal{F}$  at  $u$ . We call it the *second variation of  $\mathcal{F}$* . Moreover, again as in the elementary calculus,  $\delta^2\mathcal{F}(u)(w, w) = \frac{d^2}{dt^2}\bigg|_0 \mathcal{F}(u + tw)$ , if the second variation exists. We may summarize:

**Theorem.** *Let  $\mathcal{F} : \mathcal{S} \rightarrow \mathbb{R}$  be a functional with a local extreme in  $u \in \mathcal{S}$ . If the variation  $\delta\mathcal{F}(u)$  exists, then it has to vanish. If the second variation  $\delta^2\mathcal{F}(u)$  exists (thus in particular,  $\delta\mathcal{F}$  exists on a neighborhood of  $u$ ), then  $\delta^2\mathcal{F}(u)(w, w) \geq 0$  for a minimum, while  $\delta^2\mathcal{F}(u)(w, w) \leq 0$  for a maximum.*

**PROOF.** Assume  $\mathcal{F}$  has got a local minimum at  $u$ . We have already seen,  $f(t) = \mathcal{F}(u + tv)$  has to achieve a local minimum for each  $v$  at  $t = 0$ . Thus  $f'(0) = 0$  if  $f(t)$  is differentiable, and so  $\delta\mathcal{F}(u)$  vanishes.

Now assume  $\delta^2\mathcal{F}(u)(w, w) = f''(0) = \tau < 0$  for some  $w$ . Then the mean value theorem implies

$$f(t) - f(0) = f'(c)t = \frac{1}{c}(f'(c) - f'(0))ct$$

for some  $t \geq c > 0$ . Thus, for  $t$  small enough  $f(t) - f(0) < 0$  which contradicts  $f(0)$  being a local minimum.

The claim for maximum follows analogously (or we may apply the already proved result to the functional  $-\mathcal{F}$ ).  $\square$

<sup>8</sup>In functional analysis, this directional derivative is usually called the *Gâteaux differential*, while the continuous functional  $L$  satisfying (1) is usually called the *Fréchet differential*, going back to two of the founders of functional analysis from the beginning of the 20th century.

**Corollary.** *On top of all assumptions of the above theorem suppose  $\mathcal{F}(v + tw)$  is 2 times differentiable at  $t = 0$  and  $\delta^2 \mathcal{F}(v)(w, w) \geq 0$  for all  $v$  in a neighborhood of the critical point  $u$  and  $w \in \mathcal{S}$ . Then  $\mathcal{F}$  has got a minimum at  $u$ .*

**PROOF.** As before we consider  $f(t) = \mathcal{F}(u + tw)$ ,  $w = z - u$ . Thus, for some  $0 < c \leq 1$

$$\begin{aligned} \mathcal{F}(z) - \mathcal{F}(u) &= f(1) - f(0) = f'(0) + \frac{1}{2}f''(c) \\ &= \frac{1}{2}\delta^2 \mathcal{F}(u + cw)(w, w) \geq 0. \end{aligned}$$

□

**Remark.** Actually, the condition from the collorary is far too strong in infinite dimensional spaces. It is possible to replace it by the condition  $\delta^2 \mathcal{F}$  continuous at  $u$  and  $\delta^2 \mathcal{F}(u)(w, w) \geq C\|w\|^2$  for some real constant  $C > 0$  just in the critical point  $u$ . In the finite dimensional case, this is equivalent to the requirement  $\delta^2 \mathcal{F}$  continuous and positive definite.

**9.3.5. Back to variational problems.** As we already noticed, the answer to a variational problem minimizing a functional (we omit the arguments  $t$  of the unknown function  $u$ )

$$(1) \quad \mathcal{J}(u) = \int_{t_1}^{t_2} F(t, u, \dot{u}) dt$$

depends very much on the boundary conditions and the space of functions we deal with. If we posit  $u(t_1) = A$ ,  $u(t_2) = B$  with arbitrary  $A, B \in \mathbb{R}$  we may deal with spaces of differentiable or piecewise differentiable functions satisfying these boundary conditions. But these subspaces will not be vector spaces any more. Thus, strictly speaking, we cannot apply the concepts from the previous paragraph here.

However, we may fix any differentiable function  $v$  on  $[t_1, t_2]$  satisfying  $v(t_1) = A$ ,  $v(t_2) = B$ , e.g.  $v(t) = A + (B - A)\frac{t-t_1}{t_2-t_1}$ , and replace the functional  $\mathcal{J}$  by

$$\tilde{\mathcal{J}}(u) = \mathcal{J}(u + v) = \int_{t_1}^{t_2} F(t, u + v, \dot{u} + \dot{v}) dt.$$

Now, the intitial problem transforms to one with boundary conditions  $u(t_1) = u(t_2) = 0$  and computing the variations  $\frac{d}{d\delta} \tilde{\mathcal{J}}(u + \delta w) = \frac{d}{d\delta} \mathcal{J}(u + v + \delta w)$  does not change, i.e. we have to request  $w(t_1) = w(t_2) = 0$  and we differentiate in a vector space.

Essentially, we just exploit the natural affine structures on the subspaces of functions defined by the general boundary conditions and thus the derivatives have to live in their modeling vector subspaces.

THE FIRST AND SECOND VARIATIONS

**Corollary.** Let  $F(t, y, p)$  be a twice differentiable Lagrangian and consider the variational problem of finding a minimum of the functional (1) on the space of differentiable functions  $u \in C^1[t_1, t_2]$  with boundary conditions  $u(t_1) = A$ ,  $u(t_2) = B$ . Then the first and second variations exist and can be computed for all  $v \in \mathcal{S} = \{v \in C^1[t_1, t_2]; v(t_1) = v(t_2) = 0\}$ , as follows:

$$(2) \quad \delta\mathcal{J}(u)(v) = \int_{t_1}^{t_2} (F_y(t, u, \dot{u})v + F_p(t, u, \dot{u})\dot{v}) dt$$

$$(3) \quad \delta^2\mathcal{J}(u)(v, v) = \int_{t_1}^{t_2} (F_{yy}(t, u, \dot{u})v^2 + 2F_{yp}(t, u, \dot{u})v\dot{v} + F_{pp}(t, u, \dot{u})\dot{v}^2) dt.$$

If  $u$  is a local minimum of the variational problem, then  $\delta\mathcal{J}(u)(v) = 0$  for all  $v \in \mathcal{S}$ , while  $\delta^2\mathcal{J}(u)(v, v) \geq 0$  for all  $v$  in a neighborhood of the origin in  $\mathcal{S}$ .

**PROOF.** Thanks to our strong assumptions on the differentiability of  $F$ ,  $u$ , and  $v$ , we may differentiate the real function  $f(t) = \mathcal{J}(u + tv)$  at  $t = 0$  swapping the integral and the derivative. This immediately provides both formulae.

The remaining two claims are straightforward consequences of the theorem and corollary in the previous paragraph 9.3.4.  $\square$

**9.3.6. Euler-Lagrange equations.** We are following the path which we already tried when discussing our first bunch of examples in 9.3.3. Our next step was to guess the consequences of vanishing of the first variation in terms of a differential equation. Now we complete the arguments. We start with a simple result called the *fundamental lemma of the calculus of variation*.



**Lemma.** Assume  $u$  is a continuous function on the interval  $[t_1, t_2]$  and for all compactly supported smooth  $\varphi \in C_c^\infty[t_1, t_2]$ ,

$$\int_{t_1}^{t_2} u(t)\varphi(t) dt = 0.$$

Then  $u$  vanishes identically on  $[t_1, t_2]$ .

**PROOF.** Assume there is a  $c \in (t_1, t_2)$  such that  $u(c) > 0$ . Due to the continuity,  $u(t) > u(c)/2 > 0$  on a neighborhood  $(c - s, c + s) \subset (t_1, t_2)$ ,  $s > 0$ . Next, remind the smooth variants of indicator functions constructed in 6.1.10. For every pair of positive numbers  $0 < \varepsilon < r$ , we constructed a function  $\varphi_{\varepsilon, r}(t)$  of one real variable  $t$  such that  $\varphi_{\varepsilon, r}(t) = 1$  for  $|t| < r - \varepsilon$ , while  $\varphi_{\varepsilon, r}(t) = 0$  for  $|t| > r + \varepsilon$ , and  $0 \leq \varphi_{\varepsilon, r}(t) \leq 1$  everywhere. Thus, choosing such a function for  $\varphi$ , with  $r + \varepsilon < s$  and the origin shifted to  $t = c$ , we certainly obtain

$$\int_{t_1}^{t_2} u(t)\varphi(t) dt > \frac{1}{2}u(c)2(r - \varepsilon) > 0.$$

If we find some negative value  $u(c)$ , the same argumentation finishes the proof (notice there is no need to consider the boundary points due to the continuity of  $u$ ).  $\square$

## EULER-LAGRANGE EQUATIONS

**Theorem.** Consider a twice differentiable Lagrangian  $F(t, y, p)$  on  $[t_1, t_2] \times \mathbb{R}^2$  and a differentiable critical point  $u$  of the functional  $\mathcal{J}(u) = \int_{t_1}^{t_2} F(t, u, \dot{u}) dt$  with fixed boundary values  $u(t_1), u(t_2)$ . Then  $u$  is a solution of the differential equation

$$(1) \quad F_y(t, u, \dot{u}) - \frac{d}{dt} F_p(t, u, \dot{u}) = 0.$$

Notice that the derivative in the second term of the Euler-Lagrange equation means the so called *total derivative*, i.e. we should differentiate the composed mapping via the chain rule. This can be a problem, if we do not assume  $u$  to be twice differentiable.

**PROOF.** We already know that vanishing of the first variation  $\delta\mathcal{J}(u)$  is a necessary condition for  $u$  being a critical point. Thus we can start with the equality (2) in the previous paragraph 9.3.5 and compute by integrating per partes:

$$(2) \quad \begin{aligned} 0 &= \int_{t_1}^{t_2} (F_y(t, u, \dot{u})v + F_p(t, u, \dot{u})\dot{v}) dt \\ &= \int_{t_1}^{t_2} (F_y(t, u, \dot{u}) - \frac{d}{dt} F_p(t, u, \dot{u}))v dt \\ &\quad + F_p(t, u, \dot{u})v|_{t=t_2} - F_p(t, u, \dot{u})v|_{t=t_1}. \end{aligned}$$

Finally we exploit the above fundamental lemma of the calculus of variation with arbitrary smooth test functions  $v$  with compact supports inside  $(t_1, t_2)$ . Thus the last term with boundary values vanishes and by the lemma, the Euler-Lagrange equation has to hold true for  $u$ .  $\square$

**9.3.7. Remarks.** We made our life comfortable by taking very strong differentiability assumptions in the theorem above. In the last century, there was a lot effort to get much more general results with weaker assumptions. This is really important in practice where we need to deal with piece-wise differentiable extremals. On the other hand, we need even twice differentiable critical points in order to write down the Euler-Lagrange equation explicitly.

Another difficult point is to recognize which of the critical points are the minima or maxima of the functional. We saw that the second variation is a very specific quadratic functional, see 9.3.5(3), and there is a rich theory dealing with its properties. We do not have time to go into details here, but we mention just one simple necessary condition for the extreme (with a bit tricky proof), to get feeling about the topic).

**Lemma** (Lagrange necessary condition). Consider a twice differentiable Lagrangian  $F(t, y, p)$  on  $[t_1, t_2] \times \mathbb{R}^2$  and a differentiable critical point  $u$  of the functional  $\mathcal{J}(u) = \int_{t_1}^{t_2} F(t, u, \dot{u}) dt$  with fixed boundary values  $u(t_1), u(t_2)$ . If

$u$  is a local minimum of  $\mathcal{J}$ , then  $F_{pp}(t, u(t), \dot{u}(t)) \geq 0$  on the entire interval  $[t_1, t_2]$ .

**PROOF.** Assume there is  $t_0 \in (t_1, t_2)$  such that  $F_{pp}(t_0, u(t_0), \dot{u}(t_0)) = -\mu < 0$ . Similarly as in the proof of the fundamental lemma in the previous paragraph, we choose  $s > 0$  so that  $F_{pp}(t, u, \dot{u}) < -\frac{1}{2}\mu$  on  $(t_0 - s, t_0 + s)$ , and a smooth analog of an indicator function  $\varphi = \varphi_{\varepsilon, r}$ , centered at  $t_0$  and satisfying  $r + \varepsilon < s$ . Then the second variation evaluated on  $v(t) = \alpha\varphi(t_0 + \frac{t}{\alpha})$  with some  $\alpha > 0$  can be estimated as follows (we use a constant  $C > |F_{yy}(t, u, \dot{u})|$ ,  $C > |F_{yp}(t, u, \dot{u})|$  on the entire interval, the fact that  $\dot{v} = \dot{\varphi}$  and the derivative  $|\dot{\varphi}_{\varepsilon, r}|$  integrates over its support to two, and the substitution  $\tau = t_0 + \frac{t}{\alpha}$ )

$$\begin{aligned} \delta^2 \mathcal{J}(u)(\varphi, \varphi) &= \int_{t_1}^{t_2} (F_{yy}(t, u, \dot{u})v^2 + 2F_{yp}(t, u, \dot{u})v\dot{v} \\ &\quad + F_{pp}(t, u, \dot{u})\dot{v}^2) dt \\ &\leq \int_{t_0-\alpha s}^{t_0+\alpha s} (C\alpha^2 + 2C\alpha|\dot{v}|) dt - \frac{1}{2}\mu \int_{t_0-\alpha s}^{t_0+\alpha s} \dot{v}^2 dt \\ &= 2Cs\alpha^3 + 2C\alpha^2 \int_{t_0-s}^{t_0+s} |\dot{\varphi}| d\tau - \frac{1}{2}\mu\alpha \int_{t_0-s}^{t_0+s} \dot{\varphi}^2 d\tau \\ &= 2Cs\alpha^3 + 4Cs\alpha^2 - \frac{1}{2}\mu \left( \int_{t_0-s}^{t_0+s} \dot{\varphi}^2 d\tau \right) \alpha. \end{aligned}$$

The integral on the right-hand side is strictly positive and, thus, the entire expression is negative if  $\alpha$  is small enough. This is a contradiction and the proof is complete.  $\square$

**9.3.8. Special cases.** Very often the Lagrangians do not depend on all variables and then the variations and the Euler-Lagrange equations get special forms. The following summary is a straightforward consequence of the general equation 9.3.6(1), whose equivalent form we saw already in 9.3.3(4)



## SPECIAL FORMS OF LAGRANGIANS

**Case 1.** If the Lagrangian is  $F(t, y)$ , i.e., does not depend on the derivatives, then the Euler-Lagrange equation says

$$(1) \quad F_y(t, u) = 0$$

which is an implicit equation for  $u(t)$ . Moreover, the second variation is  $\delta^2 \mathcal{J}(u)(v, v) = \int_{t_1}^{t_2} F_{yy}(t, u)v^2 dt$ .

**Case 2.** If the Lagrangian is  $F(t, p)$ , then the Euler-Lagrange equation is

$$(2) \quad \frac{d}{dt} F_p(t, \dot{u}) = 0$$

and its solutions are given by the first order differential equation  $F_p(t, \dot{u}) = C$  with a constant parameter  $C$ . Moreover, the second variation is  $\delta^2 \mathcal{J}(u)(v, v) = \int_{t_1}^{t_2} F_{pp}(t, \dot{u})v\dot{v}^2 dt$ .

**Case 3.** If the Lagrangian is  $F(u, \dot{u})$ , then there is a consequence of the Euler-Lagrange equation (for  $\dot{u} \neq 0$ )

$$(3) \quad \frac{d}{dt} (F(u, \dot{u}) - \dot{u}F_p(u, \dot{u})) = 0$$

which again reduces the equation to the first order including a free constant parameter.

## 9.3.9. Remarks on higher dimensional problems.

**9.3.10. Problems with free boundary conditions.**

**9.3.11. Constrained and isoperimetric problems.**

placeholder

placeholder

#### 4. Complex Analytic Functions

In the rest of the chapter, we shall look at the (single complex variable) functions defined on the complex plane  $\mathbb{C} = \mathbb{R}^2$ . On many occasions we saw how helpful it was to extend objects from the real line into the complex plane. We provide a few glimpses into the rich classical theory, and we hope the readers will enjoy the use of it in the practical column.

**9.4.1. Complex derivative.** An open and connected subset  $\Omega \subset \mathbb{C}$  is called a *region*, or a *domain*. A mapping  $f : \Omega \rightarrow \mathbb{C}$  is called a *complex function of a single complex variable*. Working with complex numbers, we may repeat the definition of derivative:



##### COMPLEX DERIVATIVE

We say that a complex function  $f : \Omega \rightarrow \mathbb{C}$  has the *complex derivative*  $f'(a)$  at a point  $a \in \Omega$ , if the complex limit

$$f'(a) = \lim_{z \rightarrow a} \frac{f(z) - f(a)}{z - a} \in \mathbb{C}$$

exists. We say that  $f$  is *differentiable* in the complex sense, or *holomorphic*, on  $\Omega$ , if its complex derivative  $f'(z)$  exists at each  $z \in \Omega$ .

Clearly, this definition restricts itself to the definition of the derivative of functions of one real variable along  $\mathbb{R} \subset \mathbb{C}$ , when restricting the defining limit to real  $z$  and  $a$ . We shall see that the existence of a complex derivative is much more restrictive than in the real case.

The simplest example of a differentiable complex function is  $z \mapsto z^n$ ,  $n \in \mathbb{N}$ .

Indeed, exactly as with the real polynomials, we compute  $(z+h)^n - z^n = h(nz^{n-1} + \frac{1}{2}n(n-1)z^{n-2}h + \dots + h^{n-1})$  and thus for all  $z \in \mathbb{C}$  we obtain the limit

$$(z^n)' = \lim_{h \rightarrow 0} \frac{(z+h)^n - z^n}{h} = nz^{n-1}.$$

By the very definition, the mapping  $f \mapsto f'$  is linear over the complex scalars and thus all polynomials  $f(z)$  are differentiable this way:

$$f(z) = \sum_{k=0}^n a_k z^k \mapsto f'(z) = \sum_{k=0}^{n-1} (k+1)a_{k+1} z^k.$$

**9.4.2. Analytic functions.** A complex function  $f : \Omega \rightarrow \mathbb{C}$  is called *analytic* in the region  $\Omega$  if for each  $a \in \Omega$ , there is an open disc  $D = \{|z-a| < r\} \subset \Omega$  on which  $f$  is represented by a convergent power series

$$(1) \quad f(z) = \sum_{n=0}^{\infty} c_n (z-a)^n.$$

The restrictions of such power series to real arguments  $z$  are the analytic functions dealt with already in Chapter 5 and the direct proof of the following theorem (not relying on any real multivariate calculus) was promised already in 5.4.10.

The theorem reveals that the analytic functions behave as polynomials. The basic principle in its proof is to try to bound a series which we know very little about, by a geometric series, which we fully understand. This is a frequent technique in analysis. An important property of every power series is that it is (complex) differentiable within its radius of convergence.<sup>9</sup>

CONVERGENCE AND DERIVATIVE OF POWER SERIES

**Theorem.** Consider an analytic function (1). There exists  $R$ , which is either a non-negative real number or infinity, such that

- $f(z)$  converges absolutely if  $|z - a| < R$ ;
- $f(z)$  diverges if  $|z - a| > R$ ;
- $\frac{1}{R} = \limsup_{n \rightarrow \infty} |c_n|^{\frac{1}{n}}$ .

Further, if the radius of convergence is  $R > 0$ , then  $f$  is differentiable (in complex sense) for  $|z - a| < R$ , and its derivative equals the power series

$$f'(z) = \sum_{n=1}^{\infty} n c_n (z - a)^{n-1},$$

obtained by differentiating  $f(z)$  term by term. Moreover, the power series representing  $f'(z)$  has the same radius of convergence as  $f(z)$ .

The open disc  $|z - a| < R$  is called the *disc of convergence* of  $f(z) = \sum_{n=0}^{\infty} c_n (z - a)^n$ .

**PROOF.** Let  $L = \frac{1}{R}$ . Suppose  $|z - a| < R$ . We show that  $\sum_{n=0}^{\infty} c_n (z - a)^n$  converges. Notice, we mean that  $L = 0$  if  $R = \infty$ , and so the statement is trivially true in this case. For our fixed  $z$ , there is  $\varepsilon > 0$  such that  $(L + \varepsilon)|z - a| < 1$ .

Since  $L = \limsup_{n \rightarrow \infty} |c_n|^{\frac{1}{n}}$ , this means that  $|c_n|^{\frac{1}{n}} < L + \varepsilon$ , i.e.  $|c_n| < (L + \varepsilon)^n$  for sufficiently large  $n$ . Therefore,  $|c_n||z - a|^n \leq (L + \varepsilon)^n |z - a|^n$  and  $\sum_{n=0}^{\infty} |c_n||z - a|^n$  is majorized by the convergent geometric series  $\sum_{n=0}^{\infty} \rho^n$  with  $\rho = (L + \varepsilon)|z - a|$ . Therefore  $f(z) = \sum_{n=0}^{\infty} c_n (z - a)^n$  converges.

Now suppose  $|z - a| > R$ . By the definition of  $\limsup$ , for any  $\varepsilon > 0$ , there exists infinitely many  $c_n$  satisfying  $|c_n| > (L - \varepsilon)^n$ .

Choose  $\varepsilon > 0$  small enough such that  $(L - \varepsilon)|z - a| > 1$ . Then  $|c_n||z - a|^n > (L - \varepsilon)^n |z - a|^n$  for infinitely many  $n$ , and because  $(L - \varepsilon)|z - a| > 1$ , this implies that  $c_n (z - a)^n$  does not converge to 0 as  $n \rightarrow \infty$ . Therefore  $\sum_{n=0}^{\infty} c_n (z - a)^n$  diverges.

Next, we move to the derivative. First, notice that  $f(z) = \sum_{n=0}^{\infty} c_n (z - a)^n$  and  $g(z) = \sum_{n=1}^{\infty} n c_n (z - a)^{n-1}$  have the same radius of convergence because  $\lim n^{\frac{1}{n}} = 1$ .

<sup>9</sup>Actually the opposite implication is true as well: A holomorphic function on a domain  $\Omega$  is analytic on  $\Omega$ . We shall not provide a full proof of this result, but we come close to it below. The reader may find the full argument at nearly all basic textbooks on complex analysis.

Fix  $z_0$  in the disc of convergence, so that  $|z_0 - a| < r < R$  for some value of  $r$ . Let  $S_N(z)$ ,  $E_N(z)$  be defined by

$$S_N(z) = \sum_{n=0}^N c_n(z-a)^n, \quad E_N(z) = \sum_{n=N+1}^{\infty} c_n(z-a)^n.$$

We may think of  $S_N$ , which consists of the lower order terms of the power series, as the *main term*, and  $E_N$  as the *error term*. Notice that  $g(z) = \sum_{n=1}^{\infty} n c_n(z-a)^{n-1}$  is the term-by-term derivative of  $f(z)$ . We prove that  $f'(z_0) = g(z_0)$ , which means

$$\lim_{h \rightarrow 0} \frac{f(z_0+h) - f(z_0)}{h} - g(z_0) = 0.$$

Thus, given any  $\varepsilon > 0$ , we must show that there exists a  $\delta > 0$  such that if  $0 < |h| < \delta$ , then the expression above has absolute value less than  $\varepsilon$ . To do so, we break the expression into three parts and estimate each of those separately. More precisely, since  $f(z) = S_N(z) + E_N(z)$ , and we know the derivative  $S'_N$  of the polynomial  $S_N$ , we write

$$\begin{aligned} \frac{f(z_0+h) - f(z_0)}{h} - g(z_0) &= \\ &= \left( \frac{S_N(z_0+h) - S_N(z_0)}{h} - S'_N(z_0) \right) + \\ &+ (S'_N(z_0) - g(z_0)) + \frac{E_N(z_0+h) - E_N(z_0)}{h}. \end{aligned}$$

We analyze the individual terms. The first term contains the main term and its derivative, which exists because  $S_N$  is a polynomial. Thus, this term approaches 0 as  $h \rightarrow 0$ . In other words, given  $\frac{\varepsilon}{3} > 0$ , we can find  $\delta > 0$  such that  $0 < |h| < \delta$  implies

$$\left| \frac{S_N(z_0+h) - S_N(z_0)}{h} - S'_N(z_0) \right| < \frac{\varepsilon}{3}.$$

The second term is  $S'_N(z_0) - g(z_0)$ . Since  $S'_N(z_0) \rightarrow g(z_0)$  as  $N \rightarrow \infty$  (because we know that  $g(z)$  is a power series which converges absolutely for  $z$  in its convergence disc centered at  $a$ , and  $S'_N(z)$  is the  $N$ -th partial sum of this power series), this means that for  $\frac{\varepsilon}{3} > 0$ , we can find some  $N_1$  such that if  $N > N_1$ , then  $|S'_N(z_0) - g(z_0)| < \frac{\varepsilon}{3}$ .

The third term is the most tricky one to estimate effectively. We can write

$$E_N(z_0+h) - E_N(z_0) = \sum_{n=N+1}^{\infty} (c_n(z_0+h-a)^n - c_n(z_0-a)^n)$$

Expanding  $(z_0+h-a)^n - (z_0-a)^n = h((z_0+h-a)^{n-1} + (z_0+h-a)^{n-2}(z_0-a) + \cdots + (z_0-a)^{n-1})$ , we obtain

$$\begin{aligned} \frac{E_N(z_0+h) - E_N(z_0)}{h} &= \sum_{n=N+1}^{\infty} (c_n((z_0+h-a)^{n-1} + \\ &+ (z_0+h-a)^{n-2}(z_0-a) + \cdots + (z_0-a)^{n-1})). \end{aligned}$$

Observe, that for  $h$  sufficiently small,  $|z_0+h-a| < r$  as well as  $|z_0-a| < r$ . Therefore, if we replace all terms by their



absolute values and apply the triangle inequality, we obtain:

$$\left| \frac{E_N(z_0 + h) - E_N(z_0)}{h} \right| \leq \sum_{n=N+1}^{\infty} |c_n|nr^{n-1}.$$

The series on the right converges, and furthermore, its value approaches 0 as  $N \rightarrow \infty$ . Indeed,  $\sum_{n=N+1}^{\infty} |c_n|nr^{n-1}$  is just the tail of the series  $g(r)$  with absolute values on all of its individual terms, and we know that  $g(z)$  converges absolutely for  $|z - a| < R$ . So the series in question does converge, and since it is the end of a convergent series, its terms must approach 0 as  $N \rightarrow \infty$ . Therefore, given  $\frac{\varepsilon}{3} > 0$ , we can find  $N_2$  such that for all sufficiently small  $h$  and  $N > N_2$ ,

$$\left| \frac{E_N(z_0 + h) - E_N(z_0)}{h} \right| < \frac{\varepsilon}{3}.$$

Select now  $N > \max\{N_1, N_2\}$ . Then an application of the triangle inequality yields:

$$\begin{aligned} & \left| \frac{S_N(z_0 + h) - S_N(z_0)}{h} - S'_N(z_0) \right| + \\ & \left| S'_N(z_0) - g(z_0) \right| + \left| \frac{E_N(z_0 + h) - E_N(z_0)}{h} \right| \leq \varepsilon \quad \square \end{aligned}$$

**9.4.3. Corollaries.** We can apply the above theorem any number of times to obtain the following consequences. In particular, notice the straightforward existence of the antiderivative which we shall link with integrals in the next subsections.



COROLLARIES ON THE DERIVATIVES OF POWER SERIES

**Corollary.** Consider any power series

$$f(z) = \sum_{n=0}^{\infty} c_n(z - a)^n$$

with convergence radius  $R > 0$  and write  $D$  for its convergence disk.

(1)  $f(z)$  is infinitely (complex) differentiable in  $D$  and each of its  $k$ -th derivatives  $f^{(k)}(z)$  can be obtained by differentiating term-by-term  $k$  times. The resulting power series has radius of convergence again equal to  $R$ .

(2) There exists the (complex) antiderivative

$$F(z) = \sum_{n=0}^{\infty} \frac{1}{n+1} c_n(z - a)^{n+1},$$

such that  $F'(z) = f(z)$  in the disc of convergence  $D$ , which is the same for both series.

(3) The coefficients  $c_k$  of the power series  $f(z)$  are

$$c_k = \frac{f^{(k)}(a)}{k!}.$$

**PROOF.** All the claims are more or less obvious. Differentiating consecutively we see that  $f$  is infinitely differentiable at all  $z \in D$ , as claimed in (1). Furthermore, we see that  $f^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)c_n(z-a)^{n-k}$ , which

in particular yields (2) with  $k = 1$ . Finally, substituting  $z = a$  gives  $f^{(k)}(a) = k!c_k$ , since all terms containing  $(z - a)^{n-k}$  where  $n > k$  vanish at  $z = a$ . Therefore  $c_k = \frac{f^{(k)}(a)}{k!}$ .  $\square$

**9.4.4. Links to the real calculus.** Each complex valued function  $f$  on  $\Omega$  can be viewed as a mapping  $f : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^2$ . If this mapping is differentiable in the real sense (i.e. all partial derivatives are continuous) we may write

$$f(z) = f(a) + D^1 f(a)(z - a) + (z - a)\alpha(z)$$

for  $a \in \Omega$  and  $z$  in a small neighborhood of  $a$  in  $\Omega$ , with  $D^1 f$  being the Jacobi matrix of first partial derivatives in the two real coordinates and  $\lim_{z \rightarrow a} \alpha(z) = 0$ .

Thus it is legitimate to question whether the real linear approximation  $D^1 f(a)$  is complex linear. Obviously this happens if and only if the complex derivative  $f'(z)$  exists.

If  $f(z) = u(x + iy) + iv(x + iy)$  is the coordinate expression of a complex differentiable function  $f$  viewed as a differentiable mapping  $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,  $z = x + iy$ , then clearly

$$\frac{\partial f}{\partial x}(z) = f'(z) \cdot 1, \quad \frac{\partial f}{\partial y}(z) = f'(z) \cdot i.$$

Thus,  $\frac{\partial u}{\partial y} + i \frac{\partial v}{\partial y} = i(\frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x})$  and we have arrived at the sufficient and necessary conditions for  $D^1 f(z)$  being complex linear

$$(1) \quad u_x = v_y, \quad u_y = -v_x.$$

Yet another argument goes as follows: the rank two matrix describing multiplication by a complex number  $a + ib$  has  $a$  in the diagonal entries.  $-b$  and  $b$  are the other two entries.

In particular, this implies  $u_{xx} + u_{yy} = 0$ . The same Laplace equation holds true for the other component function  $v$  of any holomorphic function  $f = u + iv$ .

At the level of differentials, it is useful to consider two (complex valued) linear forms

$$dz = dx + i dy, \quad d\bar{z} = dx - i dy,$$

together with the dual basis at the complexified tangent space

$$\frac{\partial}{\partial z} = \frac{1}{2} \left( \frac{\partial}{\partial x} - i \frac{\partial}{\partial y} \right), \quad \frac{\partial}{\partial \bar{z}} = \frac{1}{2} \left( \frac{\partial}{\partial x} + i \frac{\partial}{\partial y} \right).$$

A straightforward check reveals that a differentiable function  $f : \Omega \subset \mathbb{C} = \mathbb{R}^2 \rightarrow \mathbb{C}$  is complex differentiable if and only if

$$\frac{\partial}{\partial \bar{z}} f = 0.$$

**9.4.5. Integrals along paths.** Another important link concerns integration along paths. A *continuous path*  $\gamma$  in the complex plane is a continuous mapping  $\gamma : J \subset \mathbb{R} \rightarrow \mathbb{C}$  defined on a bounded closed interval  $J = [a, b]$ . A path is called a *simple closed path*, or a *Jordan curve* in the complex plane<sup>10</sup>, if it does not intersect itself and  $\gamma(a) = \gamma(b)$ .

<sup>10</sup>We shall be interested in piecewise differentiable Jordan curves only, and it is quite easy to see that these curves always divide the complex planes into exactly two connected components (it is obvious for piecewise linear and the rest comes via approximation). For general Jordan curves, this is a

The composition of a path  $\gamma$  with any continuous mapping  $f$  defined on the target of  $\gamma$  is again continuous, and thus the (complex valued) Riemann integral  $\int_a^b f \circ \gamma(t) dt$  exists, but it is dependent on the parametrization of the path.

An easy way out is to restrict ourselves to differentiable paths with the derivative  $\dot{\gamma}(t) \neq 0$  for all  $t$ , and define the integral  $I_\gamma$  along a path  $\gamma$  as the Riemann integral

$$I_\gamma = \int_a^b f(\gamma(t))\dot{\gamma}(t) dt.$$

This coincides perfectly with the Riemann integrals of real functions of one variable restricted to a reparametrization  $\gamma$  of an interval in  $\mathbb{R} \subset \mathbb{C}$ .

Writing  $f = u + iv$  and  $\gamma(t) = x(t) + iy(t)$ ,

$$f(\gamma)\dot{\gamma} = (u + iv)(\dot{x} + i\dot{y}) = (u\dot{x} - v\dot{y}) + i(v\dot{x} + u\dot{y}).$$

Now we may check that the complex value  $I_\gamma$  is independent of the choice of parametrization directly by the substitution formula for real integrals. We should also notice that actually the (complex valued) linear form  $f(z) dz$  on  $\mathbb{R}^2 = \mathbb{C}$  equals

$$(u + iv)(dx + i dy) = (u dx - v dy) + i(v dx + u dy).$$

Thus,  $I_\gamma$  equals the integral of the linear form  $f(z) dz$  over the (unparametrized) submanifold  $\gamma \subset \mathbb{C}$  in the sense introduced in the first part of this chapter:

$$I_\gamma = \int_a^b (u\dot{x} - v\dot{y}) dt + i \int_a^b (v\dot{x} + u\dot{y}) dt = \int_\gamma f(z) dz.$$

In fact, any choice of parametrization ( $\dot{\gamma} \neq 0$  on  $J$ ) determines the orientation of  $\gamma$ . Thus the integral is independent of the parametrization, up to sign.

If  $\gamma$  is a composition  $\gamma_2 \circ \gamma_1$  of two paths (we simply concatenate the curves  $\gamma_1 : [a, b] \rightarrow \mathbb{C}$  and  $\gamma_2 : [b, c] \rightarrow \mathbb{C}$ ,  $\gamma_2(b) = \gamma_1(b)$ ), then clearly

$$\int_\gamma f(z) dz = \int_{\gamma_1} f(z) dz + \int_{\gamma_2} f(z) dz.$$

In particular if  $\gamma_2 = \gamma_1^{-1}$ , i.e. the same curve with opposite parametrizations, then  $\int_\gamma f(z) dz = 0$ .

Clearly, our definition of integration extends to piecewise differentiable paths. By uniform continuity over compact domains, the value  $I_\gamma$  depends continuously on the choice of the path  $\gamma$  in the  $C^0$  metric on the functions on the interval  $J$ . Thus we may approximate any integral  $I_\gamma$  by integrating the same function over a piecewise linear path  $\tilde{\gamma}$ .

**9.4.6. Antiderivatives.** If  $F(z)$  is an antiderivative of  $f(z)$ , then clearly

$$\frac{d}{dt} F(\gamma(t)) = F'(\gamma(t)) \cdot \dot{\gamma}(t) = f(\gamma(t)) \cdot \dot{\gamma}(t)$$

and therefore we have verified the straightforward generalization of the Newton integral formula in one real variable calculus:

---

difficult topological result attributed to the French mathematician Camille Jordan (1838-1922). This is the same Jordan related to the Jordan canonical form of matrices discussed in Chapter 4.

NEWTON INTEGRAL FORMULA

For each piecewise differentiable path  $\gamma : [a, b] \rightarrow \mathbb{C}$  and antiderivative  $F(z)$  of the function  $f(z)$  defined on a neighborhood of  $\gamma$ ,

$$(1) \quad I_\gamma = \int_\gamma f(z) dz = F(\gamma(b)) - F(\gamma(a)).$$

In particular, the value of the integral depends only on the values of  $F$  in the endpoints of  $\gamma$  and not on the path itself.

**Corollary.** *Antiderivatives to complex functions on connected domains are uniquely defined, up to complex constants.*

PROOF. Assume  $F'(z) = G'(z)$ , i.e.  $(F - G)'(z) = 0$ . Then for each path  $\gamma$ ,  $\gamma(0) = z$ ,  $\gamma(1) = w$ ,

$$(F - G)(w) - (F - G)(z) = \int_\gamma 0 dz = 0,$$

and thus  $F - G$  is a constant function. □

As an example, consider the paths  $\gamma_r : [0, 2\pi] \rightarrow \mathbb{C}$ ,  $\gamma_r(t) = r e^{it}$ , i.e. the positively oriented boundary of the ball  $B(0, r)$  centered at the origin, with radius  $r > 0$ . It is easy to compute the integral of  $f(z) = z^n$  along these paths, for all  $n \in \mathbb{Z}$ .

$$(2) \quad \int_{\gamma_r} z^n dz = \int_0^{2\pi} r^n e^{nit} ir e^{it} dt = \begin{cases} \frac{r^{n+1}}{n+1} [e^{(n+1)it}]_0^{2\pi} = 0 & n \neq -1 \\ ir^0 \int_0^{2\pi} e^0 dt = 2\pi i & n = -1. \end{cases}$$

In particular, we see that the integral of any polynomial along any circle vanishes (cf. more details in ??).

**9.4.7. Cauchy integral theorem.** The formula 9.4.6(1) can be applied to closed paths and we arrive immediately at the important *Cauchy integral theorem* on the convergence discs of analytic functions. This result is actually available for all holomorphic functions on much more general domains.

Recall that a domain  $\Omega$  is called *simply connected* if every simple closed continuous path in  $\Omega$  can be shrunk continuously into a point without leaving  $\Omega$ .



CAUCHY INTEGRAL THEOREM

**Theorem.** *Let  $f : \Omega \rightarrow \mathbb{C}$  be analytic on a simply connected domain  $\Omega$  and  $\gamma$  be a closed piecewise differentiable path in  $\Omega$ . Then*

$$\int_\gamma f(z) dz = 0.$$

SKETCHY PROOF. The analytic function  $f$  has an antiderivative  $F$  on each of its convergence discs. Assume first that the entire domain  $\Omega$  is contained in one such disc. Then we break the closed path  $\gamma : [0, 1] \rightarrow \Omega$  into the intervals

$0 < t_1 < t_2 < \dots < t_m = 1$  on which  $\gamma$  is differentiable, and

$$\begin{aligned} \int_{\gamma} f(z) dz &= \sum_{j=0}^{m-1} (F(\gamma(t_{j+1})) - F(\gamma(t_j))) \\ &= F(\gamma(1)) - F(\gamma(0)) = 0 \end{aligned}$$

since  $\gamma(0) = \gamma(1)$ .

In particular, if  $T$  is a triangle lying entirely in a convergence disc of the analytic function  $f(z)$  inside of  $\Omega$ , then  $\int_{\partial T} f(z) dz = 0$ , where  $\partial T$  is the oriented boundary of  $T$ .

Next, without loss of generality, the path  $\gamma$  can be viewed as a polygon, since any piecewise differentiable path  $\gamma(t)$  can be uniformly approximated by piecewise linear functions  $\gamma_n(t)$ , that form closed polygons. The integrals of  $f$  over  $\gamma_n$  approximate the integral under question. Thus, if we show  $\int_{\gamma_n} f(z) dz = 0$ , this would imply that  $\int_{\gamma} f(z) dz = 0$ , too.

It seems to be clear, that the interior of any closed polygon  $\gamma_n$  can be triangulated into closed triangles  $T_j$  so that all  $T_j$  lie in  $\Omega$  with their interiors. Actually, here we need the assumption that  $\Omega$  is simply connected if we want to fill in all the details. The integral along the path  $\gamma_n$  is then equal to the sum of the integrals over all the individual triangles (notice we integrate twice in the opposite directions over each edge which does not belong to  $\gamma_n$ ). Finally, possibly refining the polygon  $\gamma_n$  and the triangulation, we may assume that the size of each triangle  $T_j$  is so small, that  $T_j$  lies entirely in some convergence disc of  $f(z)$ . Therefore,

$$\int_{\gamma_n} f(z) dz = \sum_j \int_{\partial T_j} f(z) dz = 0,$$

and hence  $\int_{\gamma} f(z) dz = 0$  as requested.  $\square$

**9.4.8. Cauchy integral theorem again.** We were quite sloppy about the topological issues in the above sketch of the proof. Actually, there is a more general theorem deducing the conclusion of the Cauchy integral theorem under the assumption that the function  $f$  is complex differentiable (holomorphic). We shall prove this theorem under the additional assumption that  $f$  is (continuously) differentiable as a mapping of two real variables. Both conditions are obviously satisfied for analytic functions.

We remark that the general claim of the theorem is proved by a procedure similar to the above argumentation, dealing first with the claim for triangles etc. The reader may find the full proof in any basic textbook on complex analysis.

**Theorem** (Cauchy integral theorem). *If  $f(z)$  is holomorphic in a simply connected domain  $\Omega \subset \mathbb{C}$ , then for every piecewise differentiable simple closed path  $\gamma \subset \Omega$*

$$\int_{\gamma} f(z) dz = 0.$$



**PROOF OF A SPECIAL CASE.** Without loss of generality, assume that  $\gamma$  is a piecewise differentiable path bounding some simple connected region  $G$ . Write as usual  $f(z) = u(x, y) + i v(x, y)$  and so

$$\int_{\partial G} f(z) dz = \int_{\partial G} u dx - v dy + i \int_{\partial G} v dx + u dy.$$

Now, assuming  $f$  is differentiable as a functions of two real variables, the Green's version 9.1.13 of the general Stokes' theorem 9.1.12 implies

$$\begin{aligned} \int_{\partial G} f(z) dz &= \int_G \left( -\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} + i \left( \frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) \right) dx dy \\ &= 2i \int_G \frac{\partial f}{\partial \bar{z}} dx dy = 0, \end{aligned}$$

since  $\frac{\partial f}{\partial \bar{z}} = 0$  is equivalent to being holomorphic.  $\square$

The Cauchy integral theorem has an immediate consequence, ensuring the existence of antiderivatives:

**9.4.9. Theorem.** *For every analytic function  $f(z)$  in a simply connected region  $\Omega$ , its antiderivative  $F(z)$  exists in that region.*

**PROOF.** If  $\Omega$  is the convergence disc of a power series expression for  $f$ , then the claim is obvious, cf. 9.4.3.

In general, fix a point  $z_0 \in \Omega$ , and consider an arbitrary  $\zeta \in \Omega$  and any path  $\gamma \subset \Omega$  with the initial point  $z_0$  and end point  $\zeta$ . Define

$$F(\zeta) = \int_{\gamma} f(z) dz.$$

Choose any other  $\tilde{\gamma}$  with the same beginning and end, and prolong the path  $\gamma$  by  $\tilde{\gamma}^{-1}$ . This provides a closed path  $\mu = \tilde{\gamma}^{-1} \circ \gamma$ , and therefore, by the Cauchy integral theorem,  $\int_{\mu} f(z) dz = 0$ . Thus  $F(\zeta)$  is well defined, independent of the choice of  $\gamma$ .

Next, consider a small  $h$  such that the entire oriented segment  $\nu$  joining  $\zeta$  and  $\zeta + h$  is in  $\Omega$ . Then

$$F(\zeta+h) - F(\zeta) = \int_{\nu \circ \gamma} f(z) dz - \int_{\gamma} f(z) dz = \int_0^1 f(\zeta+ht) h dt.$$

In particular

$$\lim_{h \rightarrow 0} \frac{F(\zeta+h) - F(\zeta)}{h} = \lim_{h \rightarrow 0} \int_0^1 f(\zeta+ht) dt = f(\zeta)$$

and thus,  $F(\zeta)$  is the requested antiderivative.  $\square$

Clearly, the antiderivative of an analytic function on a simply connected domain is again analytic. This follows immediately from corollary 9.4.6 and the local formula for antiderivative,  $F(z) = \sum_{n=0}^{\infty} \frac{1}{n+1} c_n (z-a)^{n+1}$  of the analytic functions  $f(z) = \sum_{n=0}^{\infty} c_n (z-a)^n$  on the individual convergence discs. It is related to the much more general concept of analytic extension which we shall discuss now.

**9.4.10. Uniqueness.** At first glance, the power series representing locally an analytic function in a domain  $\Omega$  should glue together. We shall deal first with the uniqueness issues.

UNIQUENESS THEOREM

**Lemma.** Consider an analytic function  $f(z)$  in  $\Omega$  and a sequence of its zeroes  $a_n \in \Omega$  which has a limit point  $a \in \Omega$ . Then  $f(z) = 0$  everywhere in  $\Omega$ .

**PROOF.** We start with a simple observation on the non-vanishing of analytic functions:

**Claim.** Let  $f(z) \neq 0$  be analytic in  $\Omega$ , with  $f(a) = 0$  for some  $a \in \Omega$ . Then, there exists  $\varepsilon > 0$  such that  $f(z) \neq 0$  for  $0 < |z - a| < \varepsilon$ .

Indeed, in some neighbourhood of  $a$ , the analytic function  $f(z)$  is represented by a power series  $f(z) = \sum_{n=0}^{\infty} c_n(z - a)^n$ . Since  $f(a) = 0$ , we have  $c_0 = 0$ . Let  $c_k$  be the first non-zero coefficient in the series. Then  $f(z) = (z - a)^k g(z)$ , where  $g(z) = \sum_{n=k}^{\infty} c_n(z - a)^{n-k}$  and  $g(a) \neq 0$ . Therefore, by continuity of  $g(z)$ , there exists a disc centered at  $a$  of the radius  $\varepsilon > 0$  where  $g(z)$  does not have zeroes and, consequently,  $f(z) \neq 0$  for  $0 < |z - a| < \varepsilon$ .

The lemma is now a simple corollary of the above claim. Under the assumptions,  $f(z)$  must vanish identically on a non-trivial disc centered at  $a$ . Assume  $f(w) \neq 0$ ,  $w \in \Omega$ , and choose a path  $\gamma$  with  $\gamma(0) = a$ ,  $\gamma(1) = w$ . Define  $t_0$  to be the infimum of the nonempty set  $\{t \in [0, 1]; f(\gamma(t)) \neq 0\}$ . Then  $t_0 > 0$  and  $f(\gamma(t))$  is identically zero for  $t \in [0, t_0)$ . Thus the above claim applies for  $a = \gamma(t_0)$ , which leads to a contradiction with  $f(t_0) \neq 0$  and we are done.  $\square$

As a corollary, we see that any function  $f(z)$ , analytic in two concentric discs is represented in those discs by the same power series  $f(z) = \sum_{n=0}^{\infty} c_n(z - a)^n$ , where  $a$  is the centre of those discs, and  $c_k$  are Taylor coefficients  $c_k = \frac{1}{k!} f^{(k)}(a)$ ,  $k = 0, 1, \dots$

**9.4.11. Analytic extension.** The basic idea for gluing non-zero powerseries together is very simple. Consider a power series  $f(z) = \sum_{n=0}^{\infty} c_n(z - a)^n$ , converging in  $D = \{|z - a| < r\}$  and some point  $b \in D$ . If  $|z - b| + |b - a| < r$ , then  $\sum_{n=0}^{\infty} |c_n|(|z - b| + |b - a|)^n$  converges and, thus, on the smaller disc  $D_b = \{|z - b| < s = r - |b - a|\}$  we may rewrite the power series  $f(z)$  by expanding  $(z - a)^n = (z - b + b - a)^n$ :



$$\sum_{n=0}^{\infty} c_n(z - a)^n = \sum_{n=0}^{\infty} \sum_{k=0}^n \binom{n}{k} c_n (b - a)^{n-k} (z - b)^k$$

$$= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \binom{n}{k} c_n (b - a)^{n-k} (z - b)^k,$$

where all the series converge absolutely, and so the order of summation is irrelevant.

Thus, writing  $d_k = \sum_{n=k}^{\infty} \binom{n}{k} c_n (b-a)^{n-k}$ , the new power series  $f(z) = \sum_{k=0}^{\infty} d_k (z-b)^k$  converges at least on the disc  $D_b$  and we shall call it the *re-expansion* of  $\sum_{n=0}^{\infty} c_n (z-a)^n$  at the centre  $b$ .

The re-expansion is guaranteed to converge for  $|z-b| < r - |b-a|$ . However, the radius of convergence of  $\sum_{n=0}^{\infty} d_n (z-b)^n$  can be larger. The concept of analytic extension is based on this.

#### ANALYTIC ELEMENTS AND EXTENSIONS

An *analytic element*  $\Phi$  with centre at  $a \in \mathbb{C}$  is a pair  $\Phi = (D, f)$ , where  $D$  is some disc  $D = \{|z-a| \leq R\}$  and  $f$  is a convergent power series in  $D$ . The element  $\Phi$  is called *canonical*, if  $R$  is the radius of convergence of  $f$  at  $a$ .

Elements  $\Phi_1 = (D_1, f_1)$  and  $\Phi_2 = (D_2, f_2)$  are called *immediate extensions* of each other if  $D_1 \cap D_2 \neq \emptyset$  and  $f_1(z) = f_2(z)$  on  $D_1 \cap D_2$ . An element  $\Psi$  is called an *analytic extension* of  $\Phi$  along the chain  $\Phi_0 = \Phi, \Phi_1, \dots, \Phi_n = \Psi$  if  $\Phi_{j+1}$  is an immediate analytic extension of  $\Phi_j$ ,  $j = 0, 1, \dots, n-1$ .

A parametric family  $\Phi_t = (D_t, f_t)$  of canonical elements is called an *analytic extension of  $\Phi_0$  along a path*  $\gamma : [0, 1] \rightarrow \mathbb{C}$ , if (i) for all  $t \in [0, 1]$  the radii of convergence of the discs  $D_t$  are  $R_t > 0$  and the centres of  $\Phi_t$  are  $a_t = \gamma(t)$ ; and (ii) for all  $\tau \in [0, 1]$  there exist open intervals  $U_\tau \subset [0, 1]$  containing  $\tau$  such that for all  $t \in U_\tau$ ,  $\gamma(t) \in D_\tau$  and  $\Phi_t$  is an immediate extension of  $\Phi_\tau$ .

For example, the pair

$$(1) \quad \log z = \left( |z-1| < 1, \sum_{n=0}^{\infty} \frac{(-1)^n}{n} (z-1)^n \right)$$

is the canonical element restricted to the standard real logarithm function, centered at 1.

Obviously, if  $\Psi = (D_2, f_2)$  is an immediate extension of  $\Phi = (D_1, f_1)$  and the centre  $a_2$  of  $\Psi$  lies in  $D_1$  then the series  $f_2$  is just a re-expansion of  $f_1$  around  $a_2$ . This follows from the uniqueness theorem and the possibility of re-expansion.

**Lemma.** *Assume that the element  $\Phi_1 = (D_1, f_1)$  is an immediate extension of  $\Phi_0 = (D_0, f_0)$ , and that  $\Phi_2 = (D_2, f_2)$  is an immediate extension of  $\Phi_1$ . Then,  $\Phi_2$  is also an immediate extension of  $\Phi_0$  if  $D_0 \cap D_1 \cap D_2 \neq \emptyset$ .*

*If  $\Phi_t$  and  $\Psi_t$  are two analytic extensions of canonical element  $\Phi_0 = \Psi_0$  along the same path  $\gamma : [0, 1] \rightarrow \mathbb{C}$ , then  $\Psi_t = \Phi_t$ , for all  $t \in [0, 1]$ .*

**PROOF.** Clearly,  $f_0 = f_1 = f_2$  on the non-empty open subset  $D_0 \cap D_1 \cap D_2$  of  $D_0 \cap D_2$ . By the uniqueness theorem  $f_2 = f_0$  everywhere in  $D_0 \cap D_2$ , which proves the first claim.

Next, consider  $E = \{t \in [0, 1] : \Psi_t = \Phi_t\}$ . Since  $0 \in E$ , the set  $E$  is not empty. It is open, because  $U_\tau^\Phi \cap U_\tau^\Psi \subset E$  for all  $\tau \in [0, 1]$ . Further, if  $t_0 \in [0, 1]$  is a limit point of  $E$  then the elements  $\Phi_{t_0}$  and  $\Psi_{t_0}$  are immediate extensions of  $\Phi_{t_1} = \Psi_{t_1}$  for all  $t_1 \in U_{t_0}^\Phi \cap U_{t_0}^\Psi$ ,  $t_1 < t_0$ . Since  $\Phi_{t_0}$  and  $\Psi_{t_0}$  have the common centre  $\gamma(t_0)$ , they must be equal. So  $E$  is closed and thus,  $E = [0, 1]$ , i.e.  $\Phi_t = \Psi_t$  for all  $t$ .  $\square$



**9.4.12. Technical observations.** In a sequence of simple observations, we show that analytic extension along a path can be always be obtained by analytic extensions along some chain of elements and vice versa.



First, consider the family  $\Phi_t$  of elements that extend  $\Phi_0$  along some  $\gamma(t) : [0, 1] \rightarrow \mathbb{C}$  and write  $R(t)$  for the corresponding radii of convergence.

**Proposition.** *If  $R(\tau) = \infty$  for some  $\tau \in [0, 1]$ , then it is infinite for all  $t$ . If finite, then  $R(t)$  is continuous on  $[0, 1]$ .*

**PROOF.** If  $R(\tau) = \infty$ , then each element  $\Phi_t$  is a re-expansion of  $\Phi_\tau$ , and therefore  $R(t) = \infty$  for all  $t$ .

On the other hand, if  $R(\tau) < \infty$  for some  $\tau$ , then for all  $t \in U_\tau$ , the circles  $\{|z - \gamma(t)| = R(t)\}$  and  $\{|z - \gamma(\tau)| = R(\tau)\}$  intersect at a pair of points because neither of these circles lies inside the other one. Let  $w$  be one such point of intersection. Then for the triangle with vertices  $w, \gamma(\tau), \gamma(t)$  we arrive at  $|R(t) - R(\tau)| < |\gamma(t) - \gamma(\tau)|$ . Since  $\gamma$  is uniformly continuous on  $[0, 1]$ ,  $R(t)$  is continuous on  $U_\tau$ , and thus also on  $[0, 1]$ .  $\square$

**Lemma.** *Consider an analytic extension  $\{\Phi_t, t \in [0, 1]\}$  of  $\Phi_0$  along a path  $\gamma$ . There exist a finite number of intermediate points  $0 = t_0 < t_1 < \dots < t_n = 1$  such that  $\Phi_1$  is obtained by analytic extension along the chain of elements  $\Phi_0 = \Phi_{t_0}, \Phi_{t_1}, \dots, \Phi_{t_n} = \Phi_1$ .*

*Conversely, if a canonical element  $\Psi$  extends  $\Phi$  along the chain  $\Phi = \Phi_0, \Phi_1, \dots, \Phi_n = \Psi$  and  $\gamma : [0, 1] \rightarrow \mathbb{C}$  is the piece-wise linear path through the centres of  $\Phi_0, \Phi_1, \dots, \Phi_n$ , then there is a family  $\{\Psi_t\}$  of canonical elements extending  $\Psi_0 = \Phi_0$  along  $\gamma$ , such that  $\Psi_1 = \Phi_n = \Psi$ .*

**PROOF.** Since  $R(t) > 0$  for all  $t$  and is continuous, due to the compactness of  $[0, 1]$  it is separated from zero and, therefore  $R(t) > c$  for some constant  $c > 0$ . Uniform continuity of  $\gamma(t)$  implies that  $\exists \delta > 0$  such that  $|t_2 - t_1| < \delta$  yields  $|\gamma(t_2) - \gamma(t_1)| < c$ .

Since the intervals  $J_t = U_t \cap (t - \frac{\delta}{2}, t + \frac{\delta}{2})$  cover  $[0, 1]$  one can choose a finite subcover  $J_{t_1}, \dots, J_{t_{n-1}}$  for some  $t_1 < \dots < t_{n-1}$ , to which we append  $t_0 = 0$  and  $t_n = 1$  if these terminal points are missing from the sequence. Then,  $|\gamma(t_{j+1}) - \gamma(t_j)| < c$  and the centres of  $\Phi_{t_{j+1}}$  and  $\Phi_{t_j}$  lie in the disc of convergence of each other, respectively. So, these elements are re-expansions, and hence, immediate extensions of each other.

As there are a finite number of straight segments in  $\gamma(t)$ , it is sufficient to consider the case  $n = 1$ . The rest follows by induction. Thus  $\gamma(t)$  will be a segment connecting  $\gamma(0)$  and  $\gamma(1)$ , that lies in  $D_0 \cup D_1$  with  $D_0 \cap D_1 \neq \emptyset$ . Define  $\Phi_t = (D_t, f_t)$  as re-expansion of either  $\Phi_0$  or  $\Phi_1$ , depending on where  $\gamma(t)$  appears, in  $D_0$  or  $D_1$ . If  $\gamma(t) \in D_0 \cap D_1$  then re-expansion of both  $\Phi_0$  or  $\Phi_1$  at the centre  $\gamma(t)$  determine same canonical element as noticed above. The interval  $U_\tau$  can be chosen so that  $\gamma(t), t \in U_\tau$  entirely lies in either  $D_0$  or  $D_1$ . Then  $\Phi_t$  for  $t \in U_\tau$  will be an immediate extension of  $\Phi_\tau$ .  $\square$

**9.4.13. Monodromy theorem.** In general, it is very difficult to say whether there is an analytic extension of a given analytic element along a given path. But we may quite easily find out conditions under which existing extensions along paths  $\gamma_0$  and  $\gamma_1$  with a common beginning and end must coincide.



HOMOTOPIC PATHS

We say that two paths  $\gamma_0(t)$  and  $\gamma_1(t)$  with common terminal points  $\gamma_0(0) = \gamma_1(0) = a$  and  $\gamma_0(1) = \gamma_1(1) = b$  are homotopic if there exists a continuous function  $\gamma : [0, 1] \times [0, 1] \rightarrow \mathbb{C}$  such that

$\gamma(0, t) = \gamma_0(t)$ ,  $\gamma(1, t) = \gamma_1(t)$ ,  $\gamma(s, 0) = a$ ,  $\gamma(s, 1) = b$  for all  $s \in [0, 1]$ . We say that the paths  $\gamma_s(t) = \gamma(s, t)$  provide a *homotopic deformation* of  $\gamma_0$  to  $\gamma_1$ .

The following theorem says that homotopic paths lead to the same analytic extensions.

MONODROMY THEOREM

**Theorem.** *Suppose that a canonical element  $\Phi_0 = (D_0, f_0)$  centred at  $a$  can be analytically extended along every path  $\gamma_s$  in a homotopic deformation. Then the extensions along each of those paths terminate with the same canonical element  $\Phi_1$ .*

**PROOF.** Write  $\Phi_{st}$ ,  $s \in [0, 1]$ ,  $t \in [0, 1]$  the canonical elements of the extension of  $\Phi_0 = \Phi_0$  along  $\gamma_s(t)$ . Let  $R_s(t)$  be the radius of convergence of  $\Phi_{st}$ . Since  $R_s(t) > 0$  for all  $(s, t) \in [0, 1] \times [0, 1]$ , there exists  $\rho > 0$  such that  $R_s(t) > \rho$  for all  $s$  and  $t$ . Notice that  $\gamma(s, t)$  is uniformly continuous and thus, fixing  $s_0 \in [0, 1]$  we may choose an interval  $V_{s_0}$  around  $s_0$  on the  $s$ -axis such that

$$\max_{t \in [0, 1]} |\gamma(s, t) - \gamma(s_0, t)| < \frac{\rho}{4}.$$

Then, for all  $s \in V_{s_0}$ , the result of analytic extension remains the same as every element  $\Phi_{st}$  centred at  $\gamma(s, t)$  is a re-expansion of  $\Phi_{s_0t}$  because the centre  $\gamma(s, t)$  lies in the disc of convergence of  $\Phi_{s_0t}$ . Thus, extensions along  $\gamma_s(t)$  and  $\gamma_{s_0}(t)$  produce the same terminal element  $\Phi_{s_0t}$ .

Consider the set  $E = \{s \in [0, 1] : \Phi_{s1} = \Phi_{01}\}$ . Clearly,  $E$  is not empty as it contains  $s = 0$ . By the previous argument,  $E$  is open. We shall see that  $E$  is also closed. Let  $s_0$  be a limit point of  $E$ . Consider the interval  $V_{s_0}$  as constructed above. Then there is  $s' \in V_{s_0} \cap E$ , and extensions along  $\gamma_{s'}(t)$  and  $\gamma_{s_0}(t)$  coincide, and so  $s_0 \in E$ . Hence  $E = [0, 1]$ , which proves the theorem.  $\square$

**Corollary (Monodromy theorem).** *Consider a simply connected region  $\Omega \subset \mathbb{C}$  and a canonical element  $\Phi = (D, f)$  centred at  $a \in \Omega$ . Suppose that  $\Phi$  extends along any path  $\gamma \subset \Omega$  through  $a$ . Then for any  $b \in \Omega$ , the extension of  $\Phi$  along any path terminating at  $b$  is independent on the path*

and as a result, produces the same analytic element for any such path.

Thus, an analytic extension of  $\Phi$  to every point in  $\Omega$  generates an analytic function that is represented as a convergent power series in any disc inscribed in  $\Omega$ .

**9.4.14. Remarks.** We look at some simple examples where the analytic extension is crucial. Consider the function  $f(z) = \sqrt{z}$ . Clearly we may choose the two different options  $f(1) = \pm 1$  and each of the choices will lead to a canonical element. The analytic extensions of them are called the *branches* of a multivalued complex function  $f$ . Notice that  $\sqrt{z}$  is not analytic at the origin since the derivatives blow up to infinity there.

Intuitively, it seems that two closed paths in  $\mathbb{C} \setminus \{0\}$  are homotopic if and only if they run around the origin the same number of times (the *winding number*). We may imagine what happens to the values if we move  $z$  along a circle  $z = r e^{i\theta}$ . The two initial options lead to

$$f_1(z) = \sqrt{r} e^{i\theta/2} \quad f_2(z) = \sqrt{r} e^{i(\pi+\theta/2)}.$$

Once we run  $\theta$  from 0 to  $2\pi$ , the value of the branches swap. See ?? for more observations on root functions.

more in the other column

Another very important example is  $f(z) = z^{-1}$ . Since  $f(z)$  integrates to  $2\pi i$  over each circle centered in origin, there cannot exist an antiderivative to  $f$  along any of such circles. But locally, the antiderivative is the logarithmic function  $\log z$ . The canonical element 9.4.11(1) extends to one of infinite branches and running along a circle must change its value by the constant  $2\pi i$ .

We return to general analytic functions. As promised a few pages back, the monodromy theorem implies that for any analytic function  $f(z)$  in a simply connected region  $\Omega$ , there exists an analytic function representing the antiderivative. Moreover, for each analytic element of  $f$ , this is just the antiderivative of the power series representing the function on the disc. Now we may also deduce the Cauchy integral theorem for simple connected regions  $\Omega$  in another way: The integral along the closed path  $\partial\Omega$  is given by the difference of the values of the antiderivative at the terminal points. Since the boundary  $\partial\Omega$  is homotopic to a point, the integral vanishes.

**9.4.15. Cauchy theorem the third time.** The Cauchy integral theorem holds also for analytic functions on domains  $\Omega$  which are not simply connected. A bounded open domain  $\Omega \subset \mathbb{C}$  is said to have a *regular boundary*  $\partial\Omega$  if the set of boundary points  $\partial\Omega$  consists of finitely many piece-wise smooth and mutually disjoint Jordan curves  $\gamma_0, \gamma_1, \dots, \gamma_n$ .

Notice, the Jordan curves in the boundary divide the complex plane into  $n+2$  connected components. Just one of them is unbounded, one of them coincides with  $\Omega$  and all the others are bounded “holes” inside of  $\Omega$ . We write  $\gamma_0$  for the oriented exterior boundary, i.e. the boundary of the connected unbounded component of  $\mathbb{C} \setminus \Omega$  oriented counter-clockwise,



while  $\gamma_1, \dots, \gamma_n$  form the oriented interior boundary, i.e. the curves are oriented clockwise.

CAUCHY INTEGRAL THEOREM

**Theorem.** Let  $\Omega \subset \mathbb{C}$  be a bounded region with regular boundary and  $f(z)$  analytic on the closure  $\overline{\Omega}$  (i.e. analytic on some domain containing  $\overline{\Omega}$ ). Then

$$\int_{\partial\Omega} f(z) dz = 0.$$

**PROOF.** With our choice of orientation of the boundary, we must prove

$$\int_{\partial\Omega} f(z) dz = \int_{\gamma_0} f(z) dz + \sum_{j=1}^n \int_{\gamma_j} f(z) dz = 0.$$

We proceed by induction on  $n$ . If  $n = 0$ , then clearly  $\Omega$  is simply connected and the theorem is proved already, see Theorem 9.4.8.

If  $n = 1$ , then there is exactly one interior part of the boundary  $\gamma_1$  and we may choose any two smooth paths  $\mu_1$  and  $\mu_2$  joining the left most points and the right most points in  $\gamma_0$  and  $\gamma_1$ , respectively. This way, we split  $\Omega$  into two simply connected regions  $\Omega^+$  (say the upper one) and  $\Omega^-$  (the lower one), with boundaries  $\partial\Omega^+ = \gamma_0^+ \circ \mu_1 \circ \gamma_1^+ \circ \mu_2^{-1}$  and  $\partial\Omega^- = \gamma_0^- \circ \mu_2 \circ \gamma_1^- \circ \mu_1^{-1}$ . At the same time,

$$\int_{\partial\Omega} f(z) dz = \int_{\partial\Omega^+} f(z) dz + \int_{\partial\Omega^-} f(z) dz$$

since the integration over  $\mu_i$  and  $\mu_i^{-1}$ ,  $i = 1, 2$ , cancel each other on the right hand side. Moreover, the boundaries  $\partial\Omega^\pm$  are again piecewise differentiable Jordan curves and therefore both integrals on the right hand side vanish.

The general induction step is completely analogous. If  $n > 1$ , we find one of the interior boundaries  $\gamma_i$  closest to  $\gamma_0$ . Choose two cuts  $\mu_1, \mu_2$  so that one of the two newly created components of  $\Omega$  is simply connected. Then the other component has one less interior boundary and thus the theorem follows by induction. □

it needs a diagram with the cuts, see the illustration on the diagram

**9.4.16. Cauchy integral formula.** Consider an open ball without its center,  $\Omega = B(z_0, r) \setminus \{z_0\}$ , an analytic function  $f(z)$  in  $\Omega$ , and positively oriented Jordan curves  $\gamma \subset \Omega$  including  $z_0$  in its interior. Due to the Cauchy integral theorem, the integral  $\int_\gamma f(z) dz$  does depend on the choice of such  $\gamma$ . Indeed, the region enclosed between the two choices  $\gamma_1, \gamma_2$  is bounded by them, but with opposite orientations. Thus the vanishing of the integral over the boundary means that actually the integrals over  $\gamma_1$  and  $\gamma_2$  are equal.

Next, recall that the integral of  $z^{-1}$  over any circle centered at the origin is  $2\pi i$ , see 9.4.6(2). These observations suggest the following essential formula (we may expect that  $f(\zeta)$  will behave similarly as the constant  $f(z)$  for  $\gamma$  very small):



## CAUCHY INTEGRAL FORMULA

**Theorem.** Let  $f(z)$  be analytic function in the closure of the region  $\Omega \subset \mathbb{C}$  with regular boundary. Then for all  $z$  in  $\Omega$ ,

$$f(z) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(\zeta)}{\zeta - z} d\zeta.$$

**PROOF.** Fix  $z \in \Omega$  and consider an open disc  $D_\rho = \{\zeta \in \mathbb{C}; |\zeta - z| < \rho\}$  lying inside of  $\Omega$ .

The function  $g(\zeta) = \frac{f(\zeta)}{\zeta - z}$  is analytic in the closure of  $\Omega \setminus \overline{D_\rho}$ . Adopting the counterclockwise orientation of the boundary of  $D_\rho$ , the Cauchy integral theorem implies

$$\int_{\partial\Omega} \frac{f(\zeta)}{\zeta - z} d\zeta = \int_{\partial D_\rho} \frac{f(\zeta)}{\zeta - z} d\zeta.$$

We aim at showing that  $\int_{\partial D_\rho} \frac{f(\zeta)}{\zeta - z} d\zeta = 2\pi i f(z)$ . We know  $2\pi i f(z) = \int_{\partial D_\rho} \frac{f(z)}{\zeta - z} d\zeta$ . Thus, we consider

$$\int_{\partial D_\rho} \frac{f(\zeta)}{\zeta - z} d\zeta - 2\pi i f(z) = \int_{\partial D_\rho} \frac{f(\zeta) - f(z)}{\zeta - z} d\zeta$$

and estimate

$$\begin{aligned} \left| \int_{\partial D_\rho} \frac{f(\zeta) - f(z)}{\zeta - z} d\zeta \right| &\leq \max_{|\zeta - z| = \rho} \frac{2\pi\rho |f(\zeta) - f(z)|}{\rho} \\ &= 2\pi \max_{|\zeta - z| = \rho} |f(\zeta) - f(z)|. \end{aligned}$$

Clearly, the right hand side approaches zero as  $\rho \rightarrow 0$ . Hence,

$$\int_{\partial D_\rho} \frac{f(\zeta) - f(z)}{\zeta - z} d\zeta = 0$$

and the formula in the theorem has been verified.  $\square$

Notice that if we consider  $z \in \mathbb{C} \setminus \Omega$  in the above theorem, then the function  $\frac{f(\zeta)}{\zeta - z}$  is analytic in  $\Omega$  and thus the integral vanishes by the Cauchy integral theorem.

**9.4.17. Corollaries.** Taking consecutive derivatives with respect to  $z$  in the above formula, we obtain expressions for all derivatives of  $f(z)$ :

## CAUCHY INTEGRAL FORMULA FOR DERIVATIVES

**Corollary.** Let  $f(z)$  be an analytic function in the closure of the region  $\Omega \subset \mathbb{C}$  with regular boundary. Then for all  $z$  in  $\Omega$ ,

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_{\partial\Omega} \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta.$$

**PROOF.** Indeed,  $z$  is an independent argument in the smooth integrand, thus we may differentiate the integral, which yields the formula.  $\square$

Apply the Cauchy integral formula to the disc  $D_r = \{z - a < r\}$  to obtain

MEAN VALUE THEOREM

**Theorem.** For  $f(z)$  analytic in  $\overline{D_r}$  the value at the centre of the disc can be evaluated as

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + r e^{i\theta}) d\theta.$$

PROOF. By the Cauchy integral formula

$$f(a) = \frac{1}{2\pi i} \int_{|\zeta - a| = r} \frac{f(\zeta)}{\zeta - a} d\zeta.$$

Substitute  $\zeta = a + r e^{i\theta}$  and  $d\zeta = i r e^{i\theta} d\theta$  to obtain

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + r e^{i\theta}) d\theta.$$

□

**9.4.18. Laurent series.** Already in 6.3.10, we noticed that quotients  $\frac{f(z)}{g(z)}$  of two polynomials enjoy a quite nice expansion similar to power series. We called series of the form  $\sum_{n=-\infty}^{\infty} c_n (z - a)^n$  a *Laurent series*. Now we do the same with complex arguments and coefficients.

The part of the Laurent series with positive powers,  $\sum_{n=0}^{\infty} c_n (z - a)^n$ , is called the *regular part*, while the remaining part  $\sum_{n=-1}^{-\infty} c_n (z - a)^n$ , consisting of negative powers of  $(z - a)$ , is called its *principal part*. A Laurent series is called convergent if both the regular and principal parts converge.

LAURENT SERIES

**Theorem.** Every function  $f(z)$  analytic in the annulus  $A = \{r < |z - a| < R\}$ , with  $0 \leq r < R \leq \infty$  admits a representation by a Laurent series

$$(1) \quad f(z) = \sum_{n=-\infty}^{\infty} c_n (z - a)^n,$$

where the coefficients  $c_n$  can be calculated by

$$(2) \quad c_n = \int_{|\zeta - a| = \rho} \frac{f(\zeta)}{(\zeta - a)^{n+1}} d\zeta, \quad n \in \mathbb{Z}, \quad r < \rho < R.$$

The coefficients  $c_n$ , called the *Laurent coefficients* of  $f(z)$  in  $A$ , are uniquely determined. In particular they do not depend on  $\rho$ .

PROOF. Choose  $z \in A$  and  $r', R'$  such that  $r < r' < R' < R$ . The function  $f(z)$  is now analytic in the closure of the annulus  $A' = \{r' < |z - a| < R'\}$ . Therefore, by the

Cauchy integral formula, adopting the anticlockwise orientation,  $f(z)$  is a difference of two integrals,  $f(z) = J_1 - J_2$ ,

$$f(z) = \frac{1}{2\pi i} \int_{|\zeta-a|=R'} \frac{f(\zeta)}{\zeta-z} d\zeta - \frac{1}{2\pi i} \int_{|\zeta-a|=r'} \frac{f(\zeta)}{\zeta-z} d\zeta.$$

If  $|\zeta-a| = R'$ , then  $|z-a| < |\zeta-a|$  and we may expand

$$\frac{f(\zeta)}{\zeta-z} = \frac{f(\zeta)}{(\zeta-a)} \frac{1}{1 - \frac{z-a}{\zeta-a}} = \sum_{n=0}^{\infty} \frac{f(\zeta)}{(\zeta-a)^{n+1}} (z-a)^n.$$

Next, we can estimate

$$\left| \frac{f(\zeta)}{(\zeta-a)^{n+1}} (z-a)^n \right| < \frac{\max_{|\zeta-a|=R'} |f(\zeta)|}{R'} \left( \frac{|z-a|}{R'} \right)^n$$

and thus the series is uniformly convergent and admits term by term integration. Therefore

$$J_1 = \frac{1}{2\pi i} \sum_{n=0}^{\infty} c_n (z-a)^n, \quad c_n = \int_{|\zeta-a|=R'} \frac{f(\zeta)}{(\zeta-a)^{n+1}} d\zeta.$$

If  $|\zeta-a| = r'$ , then  $|\zeta-a| < |z-a|$  and similarly to the above, the expansion

$$\frac{f(\zeta)}{\zeta-z} = \frac{f(\zeta)}{(z-a)} \frac{1}{\frac{\zeta-a}{z-a} - 1} = - \sum_{n=0}^{\infty} \frac{f(\zeta)}{(z-a)^{n+1}} (\zeta-a)^n$$

leads to the equality (via term by term integration)

$$-J_2 = \frac{1}{2\pi i} \sum_{n=-1}^{-\infty} c_n (z-a)^n, \quad c_n = \int_{|\zeta-a|=r'} \frac{f(\zeta)}{(\zeta-a)^{n+1}} d\zeta.$$

Thus, we have obtained the Laurent series representation

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n$$

as requested.

On the other hand, if we are given a Laurent series (1), then fixing arbitrary  $n \in \mathbb{Z}$  and multiplying this formula by  $(z-a)^{-(n+1)}$  we can integrate over  $|z-a| = \rho$  in order to obtain

$$\int_{|z-a|=\rho} \frac{f(z)}{(z-a)^{n+1}} dz = 2\pi i c_n.$$

The circle  $\{|z-a| = \rho\}$  with  $r < \rho < R$  was chosen arbitrarily, and in particular we see that the integrals  $\int_{|z-a|=\rho} \frac{f(z)}{(z-a)^{n+1}} dz$  cannot depend on  $\rho$ .  $\square$

**9.4.19. Remarks on convergence.** Given a Laurent series, its regular part  $\sum_{n=0}^{\infty} c_n (z-a)^n$  represents a power series that converges absolutely and uniformly on compact sets in its disc of convergence  $\{|z-a| < R\}$  with  $\frac{1}{R} = \limsup_{n \rightarrow \infty} |c_n|^{\frac{1}{n}}$ , see the *Cauchy-Hadamard formula* in Theorem 9.4.2.

The principal part  $\sum_{n=-1}^{-\infty} c_n (z-a)^n$  becomes a power series  $\sum_{n=1}^{\infty} c_{-n} w^n$  after a coordinate change  $w = \frac{1}{z-a}$  and this series converges for  $|w| < \frac{1}{r}$  with  $r = \limsup_{n \rightarrow \infty} |c_{-n}|^{\frac{1}{n}}$ . Thus we have verified:

**Proposition.** For any set of coefficients  $\{c_n, n \in \mathbb{Z}\}$  set

$$\frac{1}{R} = \limsup_{n \rightarrow \infty} |c_n|^{\frac{1}{n}}, \quad r = \limsup_{n \rightarrow \infty} |c_{-n}|^{\frac{1}{n}}.$$

Then the Laurent series

$$f(z) = \sum_{n=-\infty}^{\infty} c_n(z-a)^n$$

converges absolutely and uniformly on any compact set in the annulus  $A = \{r < |z-a| < R\}$ . It is analytic in  $A$  and if  $|z-a| < r$ , then the principal part  $\sum_{n=-1}^{-\infty} c_n(z-a)^n$  diverges, while the regular part  $\sum_{n=0}^{\infty} c_n(z-a)^n$  diverges for  $|z-a| > R$ .

**9.4.20. Link to Fourier series.** There is a very interesting link between Laurent and Fourier series. If  $f$  is analytic in  $A = \{1-\rho < |z| < 1+\rho\}$  for some  $\rho > 0$  then its  $n$ -th Laurent series coefficient



$$c_n = \frac{1}{2\pi i} \int_{|z|=\rho} \frac{f(z)}{z^{n+1}} dz = \frac{1}{2\pi} \int_0^{2\pi} f(e^{it}) e^{-int} dt.$$

Therefore,  $c_n$  represents the  $n$ -th Fourier coefficient of  $\varphi(t) = f(e^{it})$  on  $t \in [0, 2\pi]$ , and the Fourier series of  $f(e^{it})$  converges uniformly to  $f(e^{it})$  on  $t \in [0, 2\pi]$ .

**9.4.21. Liouville theorem.** The formula 9.4.18(2) for Laurent series coefficients yields the following *Cauchy inequalities*:

$$(1) \quad |c_n| = \left| \frac{1}{2\pi i} \int_{|z-a|=\rho} \frac{f(z)}{(z-a)^{n+1}} dz \right| \leq \frac{\max_{|z-a|=\rho} |f(z)|}{\rho^n}$$

for all  $n \in \mathbb{Z}$ . As a straightforward consequence we obtain the following

LIOUVILLE THEOREM

**Theorem.** If  $f(z)$  is analytic in  $\mathbb{C}$  and bounded, i.e.  $|f(z)| \leq M$  for some constant  $M$  and all  $z \in \mathbb{C}$ , then  $f(z)$  is constant.

**PROOF.** The Cauchy inequalities applied to  $f(z) = \sum_{n=0}^{\infty} c_n(z-a)^n$  yield that  $|c_n| \leq \frac{M}{R^n}$ ,  $n > 0$  for any  $R > 0$ . Thus,  $c_n = 0$  for any  $n \geq 1$  and consequently  $f(z) = c_0$ . □

**9.4.22. Isolated singularities.** We look at typical examples of analytic functions around "suspicious" points. Consider the fraction  $f(z) = \frac{\sin z}{z}$ . The origin is the zero point of both  $\sin z$  and  $z$ , and since they behave very similarly for small  $z$ , we see  $\lim_{z \rightarrow 0} f(z) = 1$ .





On the other hand  $f(z) = \frac{1}{z}$  grows towards infinity,  $\lim_{z \rightarrow 0} \frac{1}{z} = \infty$ , in the sense of the *extended complex plane*  $\overline{\mathbb{C}} = \mathbb{C} \cup \infty$  (also called the *Riemann sphere*). We can imagine  $\overline{\mathbb{C}}$  as the sphere with the stereographic projection onto the plane  $\mathbb{C}$ , see the picture. Then clearly  $\lim_{z \rightarrow a} f(z) = \infty$  if and only if  $\lim_{z \rightarrow 0} |f(z)| = \infty$  in the sense of standard analysis in real variables.

picture of the Riemann sphere  $\overline{\mathbb{C}}$  missing

It might easily happen that the limit does not exist at all, see the theorem below. For example, take  $f(z) = e^{\frac{1}{z}}$  around the point  $z = 0$ . It is given by the infinite principle part of a Laurent series,  $f(z) = \sum_{n=-1}^{-\infty} \frac{1}{n!} z^n$ . In general we talk about isolated singular points:

ISOLATED SINGULARITIES

If  $f(z)$  is analytic in a punctured neighbourhood  $V = \{0 < |z - a| < \rho\}$  then  $a$  is called an *isolated singular point* for  $f(z)$ . We say that the singular point is

- *removable*, if there is a finite limit  $\lim_{z \rightarrow a} f(z) = b \in \mathbb{C}$ ;
- a *pole*, if  $\lim_{z \rightarrow a} f(z) = \infty$ ;
- an *essential singularity*, if  $\lim_{z \rightarrow a} f(z)$  does not exist in  $\overline{\mathbb{C}}$ .

A function  $f(z)$  with only isolated singularities in a domain  $\Omega \subset \mathbb{C}$  and without any essential singularities is called a *meromorphic function* in  $\Omega$ .

The function  $f(z) = \tan \frac{1}{z}$  provides an example of a non-isolated singularity at  $a = 0$ , as 0 is the limit of poles  $(\frac{\pi}{2} + n\pi)^{-1}$ ,  $n \in \mathbb{Z}$ , of  $f(z)$ .

On the other hand, all rational functions  $f(z)/g(z)$  are meromorphic in  $\mathbb{C}$ .

The following theorem classifies isolated singularities and poles in terms of Laurent series.

**Theorem.** *The following properties are equivalent:*

- *the point  $z = a$  is a removable singularity for  $f(z)$ ;*
- *$|f(z)|$  is bounded in some punctured neighbourhood  $V = \{0 < |z - a| < \rho\}$ ;*
- *the Laurent series of  $f(z)$  in  $V = \{0 < |z - a| < \rho\}$  is the Taylor series  $f(z) = \sum_{n=0}^{\infty} c_n(z - a)^n$ , i.e. the principal part vanishes;*
- *$f(a)$  can be defined so that  $f(z)$  becomes analytic in  $\{|z - a| < \rho\}$ .*

Further, the point  $z = a$  is a pole for  $f(z)$  if and only if the principal part of the Laurent series of  $f(z)$  in  $\{0 < |z - a| < \rho\}$  contains only finitely many terms, i.e.  $f(z) = \sum_{n=-N}^{\infty} c_n(z - a)^n$ ,  $n \in \mathbb{N}$ , for some integer  $N$  (the smallest  $N$  with this property is called the *order* or the *pole* at  $z = a$ ).

Finally, the Laurent series  $f(z) = \sum_{n=-\infty}^{\infty} c_n(z - a)^n$  in the punctured neighbourhood of  $a$  contains infinitely many terms with non-zero coefficients  $c_n$ ,  $n < 0$ , if and only if  $z = a$  is an *essential singularity* for  $f(z)$ .

**PROOF.** If  $|f(z)| \leq M$  for  $0 < |z - a| < \rho$ , then by the Cauchy inequalities 9.4.21(1),  $|c_{-n}| \leq M\varepsilon^n$ ,  $n > 0$ , for all  $0 < \varepsilon < \rho$ . Therefore, all coefficients with negative indices

vanish, and

$$f(z) = \sum_{n=0}^{\infty} c_n(z-a)^n.$$

Define  $f(a) = c_0$  to obtain a power series that converges in the entire disc  $\{|z-a| < \rho\}$ .

This implies the equivalence of the four conditions in the first part of the theorem.

By the definition of a pole,  $f(z) \neq 0$  in some punctured disc  $D = \{0 < |z-a| < \rho' < \rho\}$  around  $a$  as  $\lim_{z \rightarrow a} f(z) = \infty$ . Therefore  $g(z) = \frac{1}{f(z)}$  is also analytic in  $D$  and  $\lim_{z \rightarrow a} g(z) = 0$ . Hence,  $g(z)$  is analytic in  $\overline{D}$  assuming  $g(a) = 0$  and, therefore,  $g(z) = (z-a)^N h(z)$  for some integer  $n$  and analytic function  $h(z)$  with  $h(a) \neq 0$ . Thus,  $\frac{1}{h(z)}$  is also analytic on a neighborhood of  $a$  and, therefore,

$$f(z) = \frac{1}{(z-a)^N} \sum_{n=0}^{\infty} c_n(z-a)^n.$$

Conversely, if  $f(z) = \frac{1}{(z-a)^N} h(z)$ , where  $h(a) \neq 0$ , then  $\lim_{z \rightarrow a} f(z) = \infty$  and  $a$  is a pole for  $f$ .

Finally, an isolated singularity of  $f(z)$  is neither removable nor a pole if and only if the principle part of its Laurent series is infinite and this observation finishes the proof.  $\square$

**9.4.23. Some consequences.** There are several straightforward corollaries of our classification of isolated singularities. In particular, if  $\lim_{z \rightarrow a} f(z)$  does not exist, then  $f(z)$  has really chaotic behaviour:

**Theorem.** *If  $a \in \mathbb{C}$  is an essential singularity of  $f(z)$ , then for any  $w \in \overline{\mathbb{C}}$  there is a sequence  $z_n \rightarrow a$  such that  $\lim_{n \rightarrow \infty} f(z_n) = w$ .*

**PROOF.** Let  $w = \infty$ . Since the singularity  $z = a$  is not removable,  $f(z)$  cannot be bounded in any punctured neighbourhood of  $a$ . So there exists a sequence  $z_n \rightarrow a$  such that  $\lim_{n \rightarrow \infty} f(z_n) = \infty$ .

For  $w \in \mathbb{C}$ , if in any punctured neighbourhood of  $a$  there is a point  $z$  such that  $f(z) = w$ , then by making a sequence with those points we obtain a sequence  $z_n$  such that  $f(z_n) = w$ , as required. If there is a punctured neighbourhood of  $a$  where  $f(z) \neq w$ , then  $g(z) = \frac{1}{f(z)-w}$  also has an isolated singularity at  $z = a$ , which cannot be a pole or a removable one, as otherwise  $f(z) = w + \frac{1}{g(z)}$  would have a limit as  $z \rightarrow a$ . Therefore  $z = a$  is an essential singularity for  $g(z)$  and, thus, there is a sequence  $z_n \rightarrow a$  such that  $g(z_n) = \infty$ , which implies that  $\lim_{n \rightarrow \infty} f(z_n) = w$ .  $\square$

We say that  $\infty \in \overline{\mathbb{C}}$  is an *isolated singularity* of  $f(z)$  if  $f(z)$  is analytic in  $\{|z-a| > R\}$  for some  $R > 0$ . These are straightforward consequences of the Liouville theorem if  $\infty$  is the only singularity of  $f(z)$ :

**Corollary.** *If  $f(z)$  is analytic in  $\mathbb{C}$  and  $z = \infty$  is a removable singularity for  $f(z)$  then  $f(z)$  is a constant.*

*If  $f(z)$  is analytic in  $\mathbb{C}$  and  $z = \infty$  is a pole then  $f(z)$  is a polynomial  $f(z) = \sum_{j=0}^n c_j(z-a)^j$ .*

**PROOF.** The first claim is a simple reformulation of the Liouville theorem, cf. 9.4.21.

To deal with the other claim, consider  $g(w) = f(\frac{1}{w})$ . Then  $w = 0$  is a pole for  $g(w)$ . Let  $P(w) = \sum_{j=1}^N c_j w^{-j}$  be the principal part of Laurent series for  $g(w)$ . Thus,  $h(w) = g(w) - P(w)$  is analytic in  $\mathbb{C}$  with a removable singularity at  $w = 0$ . Moreover,  $\lim_{w \rightarrow \infty} h(w) = \lim_{w \rightarrow \infty} g(w) = f(0)$ . Thus,  $|h(w)|$  is bounded, and by the Liouville theorem  $h(w) = \text{const} = f(0) = c_0$ . Hence  $f(z) = g(z^{-1}) = \sum_{j=0}^N c_j z^j$ , which is a polynomial in  $z$ .  $\square$

**9.4.24. Residues.** Next we return to the Cauchy integral theorem, with our knowledge of isolated singularities.



A *residue* of an analytic function  $f(z)$  at its isolated singular point  $a \in C$  is defined as

$$\text{res}_a f = \frac{1}{2\pi i} \int_{|z|=r} f(z) dz,$$

where  $0 < r < \rho$ . Obviously, the definition does not depend on the choice of  $r$ .

#### RESIDUE THEOREM

**Theorem.** If  $f(z)$  is represented by the Laurent series  $\sum_{n=-\infty}^{\infty} c_n (z-a)^n$ , then  $\text{res}_a f = c_{-1}$ .

Further, consider a domain  $D \subset \mathbb{C}$  and a function  $f(z)$  analytic in  $\bar{D} \setminus \{a_1, \dots, a_n\}$ , where  $a_j \in D$ ,  $j = 1, \dots, n$ . Then

$$\int_{\partial D} f(z) dz = 2\pi i \sum_{j=1}^n \text{res}_{a_j} f.$$

**PROOF.** Integrating the Laurent series  $\sum_{n=-\infty}^{\infty} c_n (z-a)^n$  and using the fact that  $\int_{|z-a|=\rho} (z-a)^n dz = 0$  unless  $n = -1$ , while  $\int_{|z-a|=\rho} (z-a)^{-1} dz = 2\pi i$ , we obtain that  $\text{res}_a f = c_{-1}$ .

Next, choose such  $\rho > 0$  that open discs  $D_j = \{|z - a_j| < \rho\}$ ,  $j = 1, \dots, n$ , have pairwise empty intersections and their closures  $\bar{D}_j$  belong to  $D$ . Then the Cauchy Integral theorem 9.4.15, applied to  $D_\rho = D \setminus \bigcup_{j=1}^n D_j$  yields

$$\begin{aligned} 0 &= \int_{\partial D_\rho} f(z) dz = \int_{\partial D} f(z) dz - \sum_{j=1}^n \int_{\partial D_j} f(z) dz \\ &= \int_{\partial D} f(z) dz - \sum_{j=1}^n \text{res}_{a_j} f. \end{aligned} \quad \square$$

**9.4.25. Residues at infinity.** Recall that when integrating along the circle  $|z - a| = R$ , we always assume counterclockwise orientation on the circle. Thus we use the minus sign in the definition:

If  $f(z)$  is analytic in the closure of an exterior of a disc  $\{|z| \geq R\}$ , then

$$\text{res}_\infty f = -\frac{1}{2\pi i} \int_{|z|=R} f(z) dz.$$

In terms of the Laurent series  $f(z) = \sum_{n=-\infty}^{\infty} c_n(z-a)^n$ , valid in  $\{|z| \geq R\}$ , we have  $\operatorname{res}_{\infty} f = -c_{-1}$ .

Note that if  $f(z)$  is analytic in  $\mathbb{C} \setminus \{a_1, \dots, a_n\}$ , then

$$\operatorname{res}_{\infty} f + \sum_{j=1}^n \operatorname{res}_{a_j} f = 0.$$

Indeed, by taking a disc  $\{|z| < R\}$  of sufficiently large radius that does not contain singularities on its boundary, we conclude that

$$\frac{1}{2\pi i} \int_{|z|=R} f(z) dz = \sum_{j=1}^n \operatorname{res}_{a_j} f.$$

**9.4.26. Example of applications.** Residues of analytic functions are used for the evaluation of improper integrals in real analysis. The following lemma turns out to be very useful for such purposes. We shall write  $M(R)$  for the maximum of  $|f(z)|$  over the upper half of the circle with radius  $R$ , i.e.  $M(R) = \max_{|z|=R, \operatorname{Im} z \geq 0} |f(z)|$ .



#### JORDAN'S LEMMA

**Lemma.** Consider the function  $f(z)$  continuous on  $\{\operatorname{Im} z \geq 0, |z| = R\}$ . Then, for each positive real parameter  $t$ ,

$$\int_{|z|=R, \operatorname{Im} z \geq 0} f(z) e^{itz} dz \leq \frac{\pi}{t} M(R).$$

Consequently, if  $f(z)$  is continuous on  $\{\operatorname{Im} z \geq 0, |z| \geq R_0\}$  and  $\lim_{R \rightarrow \infty} M(R) = 0$ , then

$$\lim_{R \rightarrow \infty} \int_{|z|=R, \operatorname{Im} z \geq 0} f(z) e^{itz} dz = 0.$$

**PROOF.** We estimate the integral from the lemma:

$$\begin{aligned} \left| \int_0^{\pi} f(R e^{i\theta}) e^{-tR \sin \theta + itR \cos \theta} iR e^{i\theta} d\theta \right| \\ \leq R M(R) \int_0^{\pi} e^{-tR \sin \theta} d\theta. \end{aligned}$$

To evaluate the latter integral, we observe that  $\sin \theta \geq \frac{2}{\pi} \theta$  for  $0 \leq \theta \leq \frac{\pi}{2}$ . Thus, using  $t > 0$  and the substitution  $\tau = \frac{2Rt\theta}{\pi}$ , we arrive at

$$\begin{aligned} R M(R) \int_0^{\pi} e^{-tR \sin \theta} d\theta &= 2R M(R) \int_0^{\frac{\pi}{2}} e^{-tR \sin \theta} d\theta \\ &\leq 2R M(R) \int_0^{\frac{\pi}{2}} e^{-\frac{2tR\theta}{\pi}} d\theta = \frac{\pi}{t} M(R) \int_0^{Rt} e^{-\tau} d\tau \\ &= \frac{\pi}{t} M(R) (1 - e^{-Rt}) \leq \frac{\pi}{t} M(R). \end{aligned}$$

The consequence for  $R \rightarrow \infty$  is obvious.  $\square$

Typically, the Jordan lemma is used to compute improper integrals of real analytic (complex valued) functions  $g(x) =$

$f(x) e^{itx}$  along the entire real line (or rather the real or imaginary parts of such integrals). If the corresponding complex analytic function  $f(z)$  has only a finite number of poles  $a_k$  in the upper half plane and  $\lim_{R \rightarrow \infty} M(R) = 0$ , then we may compute the real integral

$$\int_{-\infty}^{\infty} g(x) dx = \lim_{R \rightarrow \infty} \int_{\gamma_R} g(z) dz = 2\pi i \sum_k \text{res}_{a_k} g(z),$$

where  $\gamma_R$  is the path composed by the interval  $[-R, R]$  and the half upper circle of radius  $R$ . See the diagram and the examples in the other column.

picture for the half circle area

**9.4.27. Concluding remarks.** Of course we have not touched on many important issues in this short introduction. These include the *conformal properties* of all analytic functions, i.e. they preserve all angles of curves, the richness of analytic functions which allow the mapping of any simply connected region  $\Omega$  bijectively to the unit open disc (i.e. both the map and its inverse are analytic, the *Riemann mapping theorem*). The proper setup for analytic extensions are the *Riemann surfaces* with their fascinating topological properties. Also, we only commented on the possibility of proving the Cauchy integral theorem for triangles just assuming the existence of complex derivative. The analyticity of all holomorphic functions then follows from the Cauchy integral formula. Moreover, we have not mentioned the functions of several complex variables at all!

We hope that all of these interesting issues will challenge the readers to go for further more detailed study in the relevant literature.

**G. Additional exercises to the whole chapter**

**9.G.1. Solution.**

**9.G.2. Solution.**

**9.G.3. Solution.**

**9.G.4. Solution.**

**9.G.5. Solution.**

Solutions of the exercises

**9.B.6.** The answer is  $4\pi$ .

**9.B.7.** The answer is  $36\pi$ .

**9.B.8.** The answer is  $\frac{65\pi}{24}$ .

**9.D.2.** The general solution has the form  $u = \Phi(x^2 - y^2)$ . Moreover,  $u_{(i)} = \sqrt{x^2 - y^2}$ ,  $x > y$ ,  $u_{(ii)} = y^2 - x^2$ , and the condition (iii) makes no sense.

**9.D.3.** The general solution has the form  $u = \Phi(x^2 + y^2)$ . Moreover, we see that  $u_{(i)} = x^2 + y^2$ ,  $u_{(ii)} = \frac{1}{1+x^2+y^2}$ ,  $u_{(iii)} = (x^2 + y^2 - 1)^2$  (unique for  $x^2 + y^2 > 1$ ).

**9.D.4.** The answer has the form  $u(x, y) = 2 \cos(y) \sin(x) - 1$ .

**9.D.8.** The general solution has the form  $u = K(y^2 - 2x) \cdot e^y$ .

**9.D.9.** The general solution has the form  $u(x, y) = K(y - x) \cdot e^y$ . Moreover,  $u_{(i)} = (y - x) \cdot e^{x(y-x)}$ ,  $u_{(ii)} = \frac{2}{x-y} \cdot e^{\frac{y^2-x^2}{2}}$ ,  $u_{(iii)} = (y - x)^2 \cdot e^{y(y-x)}$ .

**9.D.10.** The solution is given by  $u = \frac{y}{yC(\frac{1}{y}-x)-1}$ ,  $u_p = \frac{y}{2xy+3y-3}$ .

**9.D.11.** The solution has the form  $u = x^2 + y^2 + C(\frac{y}{x})$ ,  $u_p = x^2 + y^2 + \frac{y^2}{x^2}$ .

**9.D.12.** The solution is given by  $u = \frac{1}{4}x^2 - \frac{1}{2}y^2 + C(y\sqrt{x})$ ,  $u_p = \frac{1}{4}(x^2 - 2y^2 - xy^2)$ .

**9.D.13.** The solution is given by  $u = y^2 C\left(ye^{\frac{1}{x}}\right)$ ,  $u_p = y^3 e^{\frac{1}{x}-1}$ .

**9.D.15.** The solution is given by  $u(x, y) = \sqrt{x^2 + y^2}$ ,  $u(x, y) = 2 - \sqrt{x^2 + y^2}$ .

**9.D.16.** The solution is given by  $u(x, y) = \frac{x^2+y^2}{2}$ ,  $u(x, y) = -\frac{1}{2}\left(1 - \sqrt{x^2 + y^2}\right)^2$ .

**9.D.17.** The solution is the function  $u(x, y) = e^{\sqrt{x^2+y^2}-1}$ .

**9.D.18.** The solution is the function  $u(x, y) = \frac{1}{4}\left(x^2 + y^2 + 2\sqrt{x^2 + y^2} + 1\right)$ .

**9.D.19.** The solution is the function  $u(x, y) = -e^{1-\sqrt{x^2+y^2}}$ .

**9.D.20.** The answer is given by  $u(x, y) = x(1 - y)$ .

**9.D.21.** The answer is the function  $u(x, y) = x + y$ .

**9.D.22.** The answer is given as follows:  $u(x, y) = (2 - \sqrt{x})^2$ ,  $u = x$ .

**9.D.23.** The answer is  $u(x, y) = x^2 - y^2$ .

**9.D.24.** The answer is  $u(x, y) = y$ .

**9.D.26.**  $u(x, y) = De^{x^2 y^2}$ .

**9.D.27.**  $u(x, y) = x + y + Dxy$ .

**9.D.30.** We get  $\alpha_0 = 1$ ,  $\beta_0 = 0$ ,  $\alpha_1 = x$ ,  $\beta_1 = y$ ,  $\alpha_2 = x^2 - y^2$ ,  $\beta_2 = 2xy$ ,  $\alpha_3 = x^3 - 3xy^2$ ,  $\beta_3 = 3x^2y - y^3$ .

**9.D.34.** Parabolic equation,  $\xi = y + \ln x$ ,  $\eta = y$ ,  $u_{\eta\eta} = 0$ , solution  $u = yC(y + \ln x) + D(y + \ln x)$ .

**9.D.35.** Hyperbolic equation,  $\xi = xy$ ,  $\eta = y$ ,  $u_{\xi\eta} = 0$ ,  $u = F(y) + G(xy)$ .

**9.D.36.** Elliptic equation,  $\xi = x + 2y$ ,  $\eta = \sqrt{7}x$ ,  $\Delta u = 0$ ,  $u(x, y) = C(x + 2y + i\sqrt{7}x) + D(x + 2y - i\sqrt{7}x)$ .

**9.D.37.** Elliptic equation,  $\xi = y^2$ ,  $\eta = x^2$ ,

$$\Delta u = -\frac{1}{2}\left(\frac{u_\eta}{\eta} + \frac{u_\xi}{\xi}\right).$$

**9.D.38.** Substitute formula into equation and initial condition.

**9.D.40.**

$$u_a = \frac{1}{2} [\sin(x-t) + \sin(x+t) + (x+t)\sin(x+t) + \cos(x+t) - (x-t)\sin(x-t) - \cos(x-t)],$$

$$u_b = 2x + \frac{1}{2} [(x+t)\ln(1+(x+t)^2) - 2(x+t) + 2\arctan(x+t) - (x-t)\ln(1+(x-t)^2) + 2(x-t) - 2\arctan(x-t)],$$

$$u_c = x + \ln \sqrt{\frac{x+t}{x-t}} + \sin x - \sin x \cos t.$$

**9.D.41.**

$$u(t, x) = x^2 + 2x + t^2 + 2\sin x - 2\sin x \cos t + \frac{1}{2} [(x+t)\sin(x+t) - (x-t)\sin(x-t) + \cos(x+t) - \cos(x-t)].$$

**9.D.42.**

$$u(t, x) = 2x^2 - x + 2t^2 + 2\cos x - 2\cos x \cos t + \frac{1}{2} [-(x+t)\cos(x+t) + (x-t)\cos(x-t) + \sin(x+t) - \sin(x-t)].$$

**9.D.43.**

$$u(t, x) = x^3 + 12xt^2 + \frac{1}{4} \ln \frac{x+2t + \sqrt{1+(x+2t)^2}}{x-2t + \sqrt{1+(x-2t)^2}} + \frac{1}{4}t^2 + \frac{1}{8} \cos 2x - \frac{1}{8} \cos 2x \cos 4t.$$

**9.D.44.**

$$u(t, x) = x^2 + 9t^2 + \frac{1}{6} [(x+3t)e^{x+3t} - (x-3t)e^{x-3t} + e^{x-3t} - e^{x+3t}] + \frac{4}{9} [\cos x \cos 3t - \cos x].$$

**9.D.48.**

$$u(x, y) = \sin y \cdot \left( \frac{1}{1 - e^{2\pi}} e^x + \frac{1}{1 - e^{-2\pi}} e^{-x} \right).$$

**9.D.49.**

$$u(t, x) = -\cos \frac{\pi x}{l} \cos \frac{\pi t}{l} + \frac{l}{2\pi} \cos \frac{2\pi x}{l} \sin \frac{2\pi t}{l}.$$

**9.D.50.**

$$u(t, x) = -e^{-4t} \sin(2x) + e^{-16t} \sin(4x).$$

**9.D.51.**

$$u(t, x) = -3\sin(4x)\cos(4t) - \sin x \sin t + \sin(2x)\sin(2t)$$

**9.D.52.**

$$u(t, x) = \sum_{n=1}^{\infty} c_n e^{-\left[\frac{(2n-1)\pi}{2a}\right]^2 t} \sin\left(\frac{2n-1}{2a}\pi x\right), \quad c_n = \frac{2}{a} \int_0^a x(2a-x) \sin\left(\frac{2n-1}{2a}\pi x\right) dx = \frac{32a^2}{\pi^3(2n-1)^3}$$

**9.D.53.**

$$u(t, x) = \frac{8A^2}{\pi^3} \sum_{n=1,3,5,\dots}^{\infty} \frac{1}{n^3} \sin \frac{n\pi x}{A} \cos \frac{n\pi t}{A}.$$

**9.D.54.**

$$u(x, y) = \sum_{n=1}^{\infty} \sin \frac{n\pi x}{a} (a_n e^{\frac{n\pi}{a}y} + b_n e^{-\frac{n\pi}{a}y}), \quad a_n = \frac{1}{1 - e^{2n\pi}} \cdot \frac{2}{a} \int_0^a x(a-x) \sin \frac{n\pi x}{a} dx, \quad b_n = \frac{1}{1 - e^{-2n\pi}} \cdot \frac{2}{a} \int_0^a x(a-x) \sin \frac{n\pi x}{a} dx$$

$$u(x, y) = \sum_{n=0}^{\infty} \sin \frac{(2n+1)\pi x}{a} \left( \frac{8a^2}{(2n+1)^3\pi^3 (1 - e^{2(2n+1)\pi})} e^{\frac{(2n+1)\pi}{a}y} + \frac{8a^2}{(2n+1)^3\pi^3 (1 - e^{-2(2n+1)\pi})} e^{-\frac{(2n+1)\pi}{a}y} \right).$$



**9.D.55.**

$$u(t, x) = - \sum_{n=0}^{\infty} \frac{8}{(2n+1)^2 \pi^2} \cos\left(\frac{2n+1}{2} \pi x\right) \cdot e^{-(2n+1)^2 \pi^2 t}.$$

**9.D.58.**

$$u(x, y) = \frac{xy}{A^2}.$$

**9.D.59.**

$$u(x, y) = \frac{5}{2} + \frac{A^2}{2} \cdot \frac{x^2 - y^2}{(x^2 + y^2)^2}.$$

**9.D.60.**

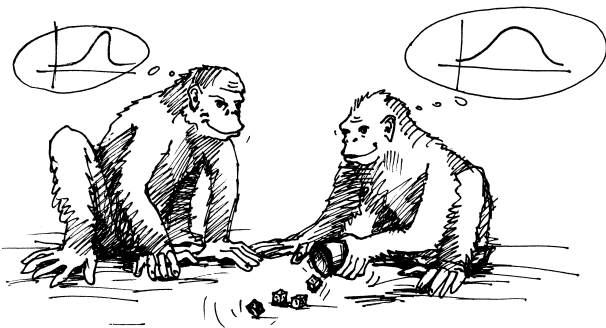
$$u(x, y) = \frac{1}{2} + \frac{x^2 - y^2}{2} + y.$$

**9.D.62.**

$$u(t, x) = \operatorname{erf}\left(\frac{x}{2\sqrt{t}}\right) - \operatorname{erf}\left(\frac{x-1}{2\sqrt{t}}\right).$$

## Statistics and probability methods

*Is statistics a part of mathematics?  
– whenever it is so, we need much of mathematics there...!*



### A. Dots, lines, rectangles

The obtained data from reality can be displayed in many ways. Let us illustrate some of them.

**10.A.1. Presenting the collected data.** 20 mathematicians were asked about the number of members of their household. The following table displays the frequency of each number of members.

Number of members	1	2	3	4	5	6
Number of households	5	5	1	6	2	1

Create the frequency distribution table. Find the mean, median and mode of the number of members. Build a column diagram of the data.

**Solution.** Let us begin with the frequency distribution table. There, we write not only the frequencies, but also the cumulative frequencies and relative frequencies (i. e., the probability that there is a given number of members in a randomly picked household). Let us denote the number of members by  $x_i$ , the corresponding frequency by  $n_i$ , the relative frequency by  $p_i$  ( $= n_i / \sum_{j=1}^6 n_j = n_i / 20$ ), the cumulative frequency by  $N_i$  ( $= \sum_{j=1}^i x_j$ ), and the relative cumulative frequency by  $F_i$

Roughly speaking, statistics is any processing of numerical or other type of data about a population of objects and their presentation. In this context, we talk about *descriptive statistics*. Its objective is thus to process and comprehensibly represent data about objects of a given “population” — for instance, the annual income of all citizens obtained from the complete data of revenue authorities, or the quality of hotel accommodation in some region. In order to achieve this, we focus on simple numerical characterization and visualization of the data.

in general, many pictures missing!!

Mathematical statistics uses mathematical methods to derive conclusions valid for the whole (potentially infinite) population of objects, based on a “small” sample. For instance, we might want to find out how much a certain disease is spread in the population by collecting data about a few randomly chosen people, but we interpret the results with regard to the entire population. In other words, mathematical statistics makes conclusions about a large population of objects based on the study of a small (usually randomly selected) sample collection. It also estimates the reliability of the resulting conclusions.

Mathematical statistics is based on the tools of probability theory, which is very useful (and amazing) in itself. Therefore, probability theory is discussed first.

This chapter provides an elementary introduction to the methods of probability theory, which should be sufficient for correct comprehension of ordinary statistical information all around us. However, for a serious understanding of a mathematical statistician’s work, one must look for other resources.

### 1. Descriptive statistics

Descriptive statistics alone is not a mathematical discipline although it uses many manipulations with numbers and sometimes even very sophisticated methods. However, it is a good opportunity for illustrating the mathematical approach to building generally useful tools.

At the same time, it should serve as a motivation for studying probability theory because of later applications in statistics.

$$(\text{= } N_i/20 = \sum_{j=1}^i p_j):$$

$x_i$	$n_i$	$p_i$	$N_i$	$F_i$
1	5	1/2	5	1/4
2	5	1/4	10	1/2
3	1	1/20	11	11/20
4	6	3/10	17	17/20
5	2	1/10	19	19/20
6	1	1/20	20	1

Now, we can easily construct the wanted (column) graphs of (relative, cumulative) frequencies:

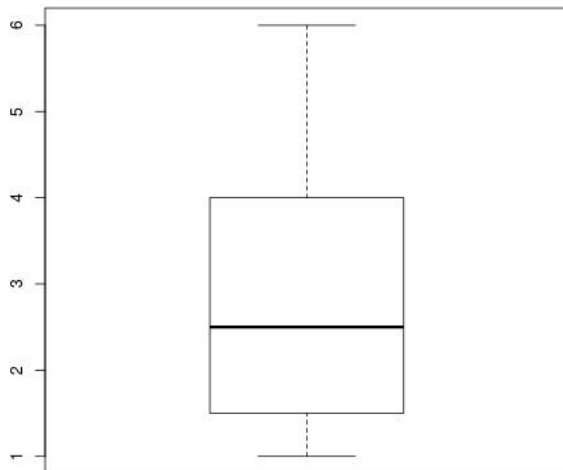
The mean number of members of a household is:

$$\bar{x} = \frac{5 \cdot 1 + 5 \cdot 2 + 1 \cdot 3 + 6 \cdot 4 + 2 \cdot 5 + 1 \cdot 6}{20} = 2.9.$$

The median is the arithmetic mean of the tenth and eleventh values (having been sorted), which are respectively 2 and 3, i. e.,  $\tilde{x} = 2.5$ .

The mode is the most frequent value, i. e.,  $\hat{x} = 4$ .

The collected data can also be presented using a *box plot*:



The upper and lower sides of the “box” correspond respectively to the first (lower) and the third (upper) quartile, so its height is equal to the interquartile range. The thick horizontal line is drawn at the median level; the lower and upper horizontal lines correspond respectively to the minimum and maximum elements of the data set, or to the value that is 1.5 times the interquartile range less than the lower side of the box (and greater than the upper side, respectively). The data outside this range would be shown as circles.

We can also build the histogram of the data:

In our brief introduction, we first introduce the concepts allowing to measure the positions of data values and the variability of the data values (means, percentiles etc.). We touch the problem how to visualize or otherwise present the data sets (diagrams). Then we deal with the potential relations between more data sets (covariance and principal components) and, finally, we deal with data without numerical values relying just on their frequencies of appearance (entropy).

### 10.1.1. Probability, or statistics?



It is not by accident that we return to a part of the motivating hints from the first chapter, as soon as we have managed to gather enough mathematical tools both discrete and continuous.

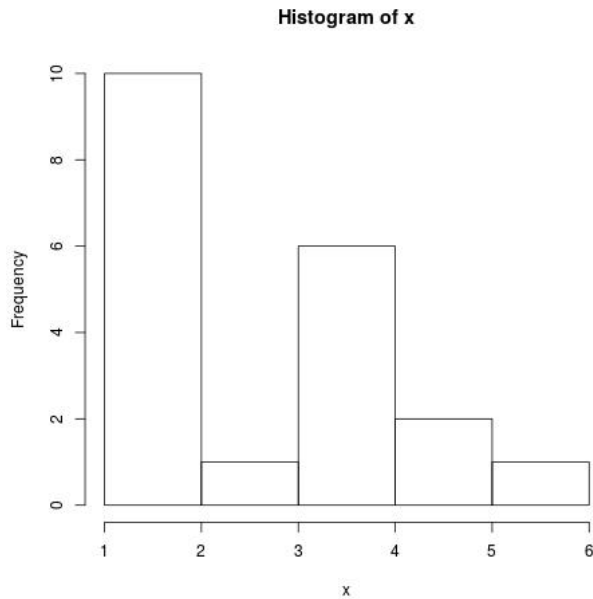
Nowadays, many communications are of a statistical nature, be it in media, politics, or science. Nevertheless, in order to properly understand the meaning of such a communication and using particular statistical methods and concepts, one must have a broad knowledge of miscellaneous parts of mathematics. In this subsection, we move away from the mathematical theory; and think about the following steps and our objectives.

As an example of a population of objects, consider the students of a given basic course. Then, the examined numerical data can be:

- the “mean number of points” obtained during the course in the previous semester and the “variance” of these values,
- the “mean marks” for the examination of this and other courses and the “correlation” (i.e. mutual dependence) of these results,
- the “correlation” of data about the past results of given students,
- the “correlation” of the number of failed exams of a given student and the number of hours spent in a temporary job,
- ...

With regard to the first item, the arithmetic mean itself does not carry enough information about the quality of the lecture or of the lecturer, nor about the results of particular students. Maybe the value which is “in the middle” of the population, or the number of points achieved by the student who was just better than half of the students is of more concern. Similarly, the first quarter, the last quarter, the first tenth, etc. maybe of interest. Such data are called *statistics* of the population. Such statistics are interesting for the students in question as well, and it is quite easy to define, compute, and communicate them.

From general experience or as a theoretical result outside mathematics, a reasonable assessment should be “normally” distributed. This is a concept of probability theory, and it requires quite advanced mathematics to be properly defined. Comparing the collected data about even a small random population of students to theoretical results can serve in two ways: We can estimate the parameters of the distribution as well as draw a conclusion whether the assessment is reasonable.



Note that the frequencies of one- and two-member households were merged into a single rectangle. This is used in order to make the data “easier to read” – there exist (various and ambiguous) rules for the merging.

We simply mention this fact without presenting an exact procedure (it is just as anyone likes). □

**10.A.2.** Given a data set  $x = (x_1, x_2, \dots, x_n)$ , find the mean and variance of the centered values  $x_i - \bar{x}$  and the standardized values  $\frac{x_i - \bar{x}}{s_x}$ .

**Solution.** The mean of the centered values can be found directly using the definition of arithmetic mean:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{\bar{x}}{n} \sum_{i=1}^n 1 = \bar{x} - \bar{x} = 0.$$

The variance of the centered values is clearly the same as for the original ones ( $s_x$ ). For the standardized values, the mean is equal to zero again, and the variance is

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right)^2 = \frac{1}{s_x^2} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = 1. \quad \square$$

**10.A.3.** Prove that the variance satisfies  $s_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$ .

**Solution.** Using the definitions of variance and arithmetic mean, we get:

$$\begin{aligned} s_x^2 &= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{2\bar{x}}{n} \sum_{i=1}^n x_i + \bar{x}^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2. \end{aligned}$$

At the same time, the numerical values of statistics for a given population can yield qualitative description of the likelihood of our conclusions. We can compute statistics which reflect the variability of the examined values, rather than where these values are positioned within a given population. For instance, if the assessment does not show enough variability, it may be concluded that it is badly designed, because the students’ skills are of course different. The same applies if the collected data seem completely random.

In the above paragraph, it is assumed that the examined data is reliable. This is not always the case in practice. On the contrary, the data is often perturbed with errors due to construction of the experiment and the data collection itself.

In many cases, not much is known about the type of the data distribution. Then, methods of non-parametric statistics are often used (to be mentioned at the end of this chapter). Very interesting conclusions can be found if we compare the statistics for different quantities and then derive information about their relations. For example, if there is no evident relation between the history of previous studies and the results in a given course, then it may be that the course is managed wrongly.

These ideas can be summarized as follows:

- In descriptive statistics, there are tools which allow the understanding of the structure and nature of even a huge collection of data;
- in mathematics, one works with an abstract mathematical description of probability, which can be used for analysis of given data. Especially, this is when there is a theoretical model to which the data should correspond;
- conclusions of statistical investigation of samples of particular data sets can be given by mathematical statistics;
- mathematical statistics can also estimate how adequate such a description is for a given data set.

**10.1.2. Terminology.** Statisticians have introduced a great many concepts which need mastering. The fundamental concept is that of a *statistical population*, which is an exactly defined set of basic *statistical units*. These can be given by enumeration or by some rules, in case of a larger population.



On every statistical unit, *statistical data* is measured, with the “measurement” perceived very broadly.

For instance, the population can consist of all students of a given university. Then, each of the students is a *statistical unit* and much data can be gathered about these units – the numerical values obtainable from the information system, what is their favorite colour, what they had for dinner before their fast test, etc.

The basic object for examining particular pieces of data is a *data set*. It usually consists of ordered values. The ordering can be either natural (when the data values are real numbers, for example) or we can define it (for instance, when we observe colours, we can express them in the RGB format

**10.A.4.** The following values have been collected:

10; 7; 7; 8; 8; 9; 10; 9; 4; 9; 10; 9; 11; 9; 7; 8; 3; 9; 8; 7.

Find the arithmetic mean, median, quartiles, variance, and the corresponding box diagram.

**Solution.** Denoting the individual values by  $a_i$  and their frequencies by  $n_i$ , we can arrange the given data set into the following table.

$a_i$	3	4	7	8	9	10	11
$n_i$	1	1	4	4	6	3	1

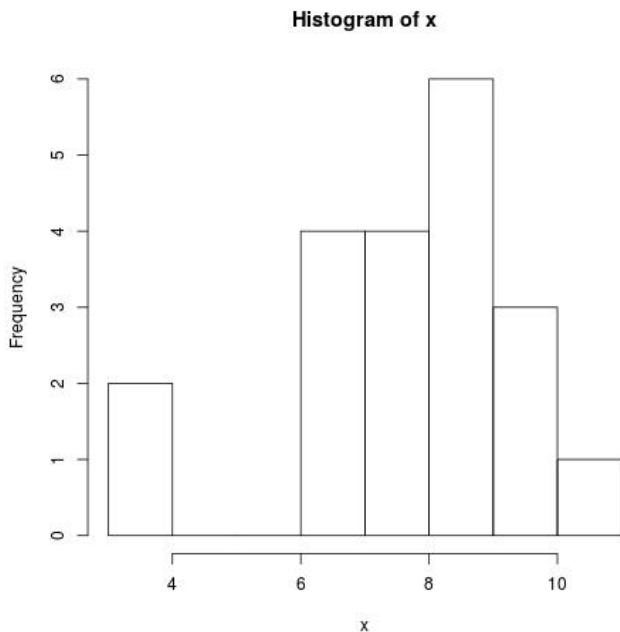
From the definition of arithmetic mean, we have

$$\bar{x} = \frac{3 + 4 + 4 \cdot 7 + 4 \cdot 8 + 6 \cdot 9 + 3 \cdot 10 + 11}{1 + 1 + 4 + 4 + 6 + 3 + 1} = \frac{162}{20} = 8.1.$$

Since the tenth least collected value is  $x_{(10)} = 8$  and the eleventh one is  $x_{(11)} = 9$ , the median is equal to  $\tilde{x} = \frac{8+9}{2} = 8.5$ . The first quartile is  $x_{0.25} = \frac{x_{(5)}+x_{(6)}}{2} = 7$ , and the third quartile is  $x_{0.75} = \frac{x_{(15)}+x_{(16)}}{2} = 9$ . From the definition of variance, we get  $s_x^2$ :

$$\frac{5 \cdot 1^2 + 4 \cdot 1^2 + 4 \cdot 1.1^2 + 4 \cdot 0.1^2 + 6 \cdot 0.9^2 + 3 \cdot 1.9^2 + 2 \cdot 9^2}{1 + 1 + 4 + 4 + 6 + 3 + 1} = 3.59.$$

The histogram and box diagram are shown in the following pictures.



where we have used "statistics" method to make the histogram "nice" and "clear". You can find a lot of these conventions in the books on statistics, but if you do not know them, you are lost. This is the default setting of the R program. For example if you replace just value 3 by 2 you get quite different looking histogram:

and order them with respect to this sign). We can also work with unordered values.

Since statistical description aims at telling comprehensible information about the entire population, we should be able to compare and take ratios of the data values. Therefore, we need to have a *measurement scale* at our disposal. In most cases, the data values are expressed as numbers. However, the meaning of the data can be quantified variously, and thus we distinguish between the following *types of data measurement scales*.

TYPES OF DATA MEASUREMENT SCALES

The data values are called:

- *nominal* if there is no relation between particular values; they are just qualitative names, i.e. possible values (for instance, political parties or lecturers at a university when surveying how popular they are);
- *ordinal* the same as above, but with an ordering (for example, number of stars for hotels in guidebooks);
- *interval* if the values are numbers which serve for comparisons but do not correspond to any absolute value (for example, when expressing temperature in Celsius or Fahrenheit degrees, the position of zero is only conventional);
- *ratio* if the scale and the position of zero are fixed (most physical and economical quantities).

With nominal types, we can interpret only equalities  $x_1 = x_2$ ; with ordinal types, we can also interpret inequalities  $x_1 < x_2$  (or  $x_1 > x_2$ ); with interval types, we can also interpret differences  $x_1 - x_2$ . Finally, with rational types, we have also ratios  $x_1/x_2$  available.

**10.1.3. Data sorting.** In this subsection, we work with a *data set*  $x_1, x_2, \dots, x_n$ , which can be ordered (thus, their type is not nominal) and which have been obtained through measurement on  $n$  statistical units. These values are sorted in a *sorted data set*

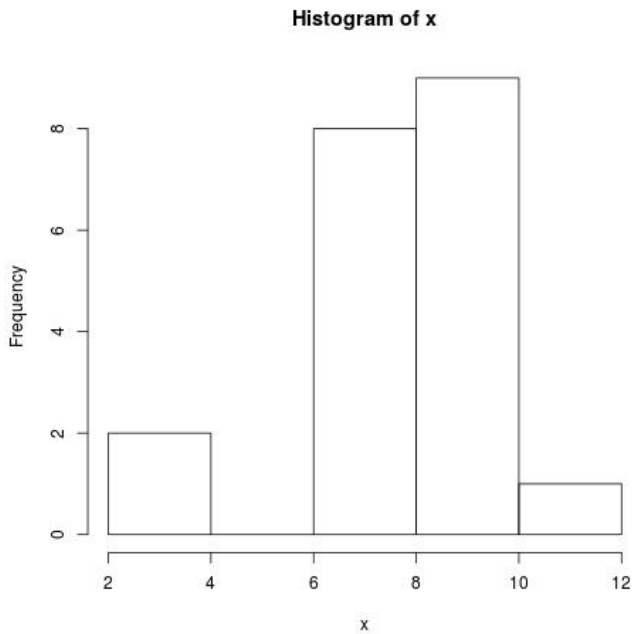
$$(1) \quad x_{(1)}, x_{(2)}, \dots, x_{(n)}.$$

The integer  $n$  is called the *size of the data set*.

When working with large data sets where only a few values occur, the simplest way to represent the data set is to enumerate the values' frequencies.

For instance, when surveying the political party preference or when presenting the quality of a hotel, write only the number of occurrences of each value.

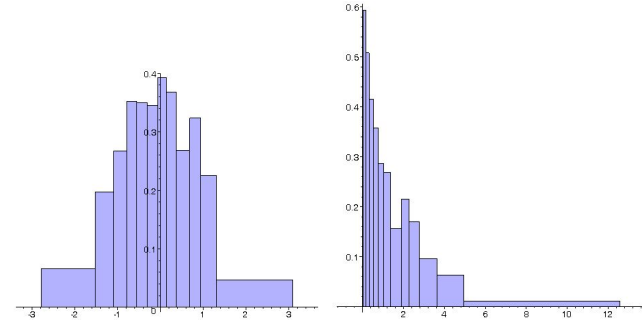
If there are many possible values (or there can even be continuously distributed real values), divide them into a suitable number of intervals and then observe the frequencies in the given intervals. The intervals are also called *classes* and the frequencies are called *class frequencies*. We also use *cumulative frequencies* and *cumulative class frequencies* which correspond to the sum of frequencies of values not exceeding a given one.



Most often, the mean  $a_i$  of a given class is considered to be its representative, and the value  $a_i n_i$  (where  $n_i$  is the frequency of the class) is the total contribution of the class. *Relative frequencies*  $a_i/n$ , and *relative cumulative frequencies*, can also be considered.

A graph which has the intervals of particular classes on one axis and rectangles above them with height corresponding to the frequency is called a *histogram*. Cumulative frequency is represented similarly.

The following diagram shows histograms of data sets of size  $n = 500$  which were randomly generated with various standard distributions (called normal,  $\chi^2$ , respectively).

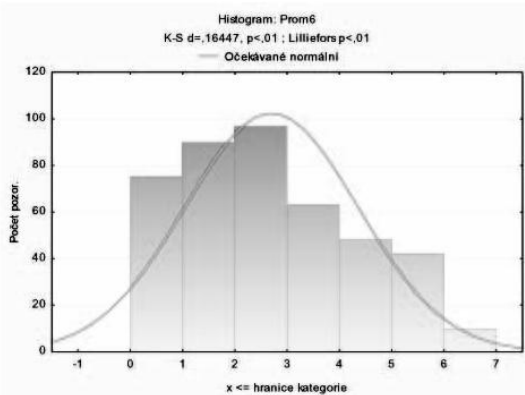


**10.A.5.** 425 carps were fished, and each one was taken weighed. Then, mass intervals were set, resulting in the following frequency distribution table:

Weight (kg)	0–1	1–2	2–3	3–4	4–5	5–6	6–7
Class midpoint	0.5	1.5	2.5	3.5	4.5	5.5	6.5
Frequency	75	90	97	63	48	42	10

Draw a histogram, find the arithmetic, geometric, and harmonic means of the carps' weights. Furthermore, find the median, quartiles, mode, variance, standard deviation, coefficient of variation, and draw a box plot.

**Solution.** The histogram looks as follows:



From the definitions of the corresponding concepts in subsection 10.1.4, we can directly compute that the arithmetic mean is  $\bar{x} = 2.7$  kg, the geometric mean is  $\bar{x}^G = 2.1$  kg, and the harmonic mean is  $\bar{x}^H = 1.5$  kg. By the definitions of subsection 10.1.5, the median is equal to  $\tilde{x} = x_{0.5} = 2.5$

**10.1.4. Measures of the position of statistical values.** If



the magnitude of values around which the collected data values gather are to be expressed, then the concepts of the definition below can be used. There, we work with ratios or interval types of scales.

Consider an (unsorted) data set  $(x_1, \dots, x_n)$  of the values for all examined statistical units and let  $n_1, \dots, n_m$  be the class frequencies of  $m$  distinct values  $a_1, \dots, a_m$  that occur in this set.

MEANS

**Definition.** The *arithmetic mean* (often only *mean*) is given as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{j=1}^m n_j a_j.$$

The *geometric mean* is given as

$$\bar{x}^G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

and makes sense for positive values  $x_i$  only. The *harmonic mean* is given as

$$\bar{x}^H = \left( \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right)^{-1}$$

and is also used for positive values  $x_i$  only.

The arithmetic mean is the only one of the three above which is invariant with respect to affine transformations. For all scalars  $a, b$ ,

$$\overline{(a + b \cdot x)} = \frac{1}{n} \sum_{i=1}^n (a + b x_i) = a + b \sum_{i=1}^n x_i = a + b \cdot \bar{x}.$$

kg, the lower quartile to  $x_{0.25} = 1.5$  kg, the upper quartile to  $x_{0.75} = 3.5$  kg, and the mode is  $\hat{x} = 2.5$  kg. From the definitions of subsection 10.1.6, we compute the variance of the weights, which is  $s_x^2 = 2.7 \text{ kg}^2$ , whence it follows that the standard deviation is  $s_x = 1.7$  kg, and the coefficient of variation is  $V_x = 0,6$ .  $\square$

**10.A.6.** Prove that the entropy is maximal if the nominal values are distributed uniformly, i. e., the frequency of each class is  $n_i = 1$ .

**Solution.** By the definition of entropy (see 10.1.11), we are looking for the maximum of the function  $H_X = -\sum_{i=1}^n p_i \ln p_i$  with respect to unknown relative frequencies  $p_i = \frac{n_i}{n}$ , which satisfy  $\sum_{i=1}^n p_i = 1$ . Therefore, this is a typical example of finding constrained extrema, which can be solved using Lagrange multipliers. The corresponding Lagrange function is

$$L(p_1, \dots, p_n, \lambda) = -\sum_{i=1}^n p_i \ln p_i + \lambda \left( \sum_{i=1}^n p_i - 1 \right).$$

The partial derivatives are  $\frac{\partial L}{\partial p_i} = -\ln p_i - 1 + \lambda$ , hence its stationary points is determined by the equations  $p_i = e^{\lambda-1}$  for all  $i = 1, \dots, n$ . Moreover, we know that the sum of the relative frequencies  $p_i$  is equal to one. This means that  $ne^{\lambda-1} = 1$ , whence we get  $\lambda = 1 - \ln n$ . Substitution then yields  $p_i = \frac{1}{n}$ .  $\square$

**10.A.7.** The following graphs depict the frequencies of particular amounts of points obtained by students of the MB104 lecture at the Faculty of Informatics of Masaryk University in 2012. The axes of the cumulative graph are “swapped”, as opposed to the previous example.

The frequencies of particular amounts of points are enumerated in the following table:

Therefore, the arithmetic mean is especially suitable for interval types.

The logarithm of the geometric mean is the arithmetic mean of the logarithms of the values. It is especially suitable for those quantities which cumulate multiplicatively, e. g. interests. If the interest rate for each time period is  $x_i\%$ , then the final result is the same as if the interest rate had constant value of  $\bar{x}^G\%$ . See 10.A.9 for an example where the harmonic mean is appropriate.

In subsection 8.1.30 (page 746), we use the methods invented there to prove that the geometric mean never exceeds the arithmetic mean. The harmonic mean never exceeds the geometric mean, and so

$$\bar{x}^H \leq \bar{x}^G \leq \bar{x}.$$

**10.1.5. Median, quartile, decile, percentile, ...** Another way of expressing the position or distribution of the values is to find, for a number  $\alpha$  between zero and one, such a value  $x_\alpha$  that  $100\alpha\%$  of values from the set are at most  $x_\alpha$  and the remaining ones are greater than  $x_\alpha$ . If such a value is not unique, one can choose the mean of the two nearest possibilities.

The number  $x_\alpha$  is called the  $\alpha$ -quantile. Thus, if the result of a contestant puts him into  $x_{1.00}$ , it does not mean that he is better than anyone else yet. However, there is surely no one better than him.

The most common values of  $x_\alpha$  are the following:

- The *median* (also sample median) is defined by

$$\tilde{x} = x_{0.50} = \begin{cases} x_{((n+1)/2)} & \text{for odd } n \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{for even } n \end{cases},$$

where  $x_{(k)}$  corresponds to the value in the sorted data set 10.1.3(1).

- The *first and third quartile* are  $Q_1 = x_{0.25}$  and  $Q_3 = x_{0.75}$ , respectively.
- The *p-th quantile* (also *sample quantile* or *percentile*)  $x_p$ , where  $0 < p < 1$  (usually rounded to two decimal places).

One can also meet the *mode*, which is the value  $\hat{x}$  that is most frequent in the data set  $x$ .

The arithmetic mean, median (with ratio types), and mode (with ordinal or nominal types) correspond to the “anticipated” values.

Note that all  $\alpha$ -quantiles with interval scales are invariant with respect to affine transformations of the values (check this yourselves!).

**10.1.6. Measures of the variability.** Surely any measure of the variability of a data set  $x \in \mathbb{R}^n$  should be invariant with respect to constant translations. In the Euclidean space  $\mathbb{R}^n$ , both the standard distance and the sample mean have this property.

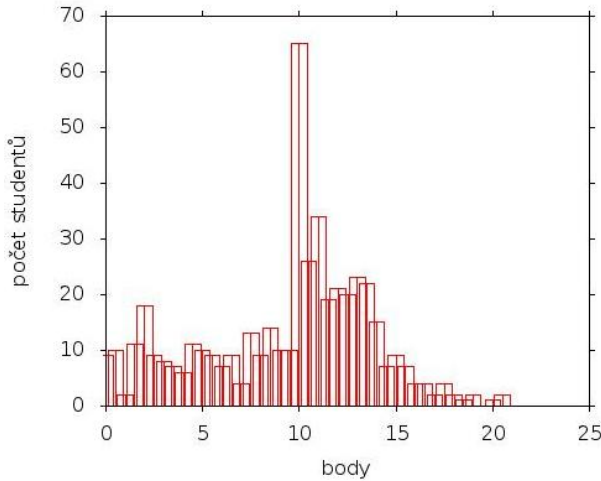
Therefore, choose the following:



# of points	# of students
20.5	1
20	1
19	2
18.5	1
18	2
17.5	3
17	2
16.5	4
16	3
15.5	5
15	7
14.5	6
14	14
13.5	21
13	21
12.5	19
12	17
11.5	18
11	31
10.5	22
10	53

# of points	# of students
9.5	9
9	9
8.5	13
8	8
7.5	13
7	4
6.5	7
6	4
5.5	8
5	7
4.5	9
4	5
3.5	7
3	8
2.5	8
2	14
1.5	8
1	2
0.5	6
0	9

The corresponding histogram looks as follows:



The histogram was obtained from the Information System of Masaryk University. We can see that the data are shown in a somewhat unusual way: individual amounts of points correspond to “double rectangles”. It is a matter of taste how to represent the data (it is possible to merge some values, thereby decreasing the number of rectangles, or to use thinner rectangles).

VARIANCE AND STANDARD DEVIATION

**Definition.** The *variance* of a data set  $x$  is defined by

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The *standard deviation*  $s_x$  is defined to be the square root of the variance.

As requested, the variability of statistical values is independent of constant translation of all values. Indeed, the unsorted data set

$$y = (x_1 + c, x_2 + c, \dots, x_n + c)$$

has the same variance  $s_y = s_x$ .

Sometimes, the *sample variance* is used, where there is  $(n - 1)$  in the denominator instead of  $n$ . The reason will be clear later, cf. 10.3.2.

In case of class frequencies  $n_j$  of values  $a_j$  for  $m$  classes, this expression leads to the value

$$s_x^2 = \frac{1}{n} \sum_{j=1}^m n_j (a_j - \bar{x})^2$$

of the variance. In practice, it is recommended to use the Sheppard’s correction, which decreases  $s_x^2$  by  $h^2/12$ , where  $h$  is the width of the intervals that define the classes.

Further, one can encounter the *data-set range*

$$R = x_{(n)} - x_{(1)}$$

and the *interquartile range*

$$Q = Q_3 - Q_1.$$

The *mean deviation*, which is defined as the mean distance of the values from the median:

$$D_x = \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}|.$$

The following theorem clarifies why these measures of variability are chosen:

**Theorem.** The function  $S(t) = (1/n) \sum_{i=1}^n (x_i - t)^2$  has the minimum value at  $t = \bar{x}$ , i.e., at the sample mean.

The function  $D(t) = (1/n) \sum_{i=1}^n |x_i - t|$  has the minimum value at  $t = \tilde{x}$ , i.e., the median.

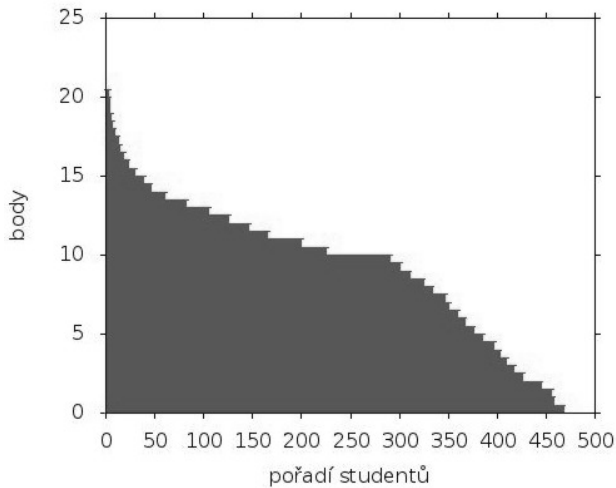
**PROOF.** The minimum of the quadratic polynomial  $f(t) = \sum_{i=1}^n (x_i - t)^2$  is at the only root of its derivative:

$$f'(t) = -2 \sum_{i=1}^n (x_i - t).$$

Since the sum of the distances of all values from the sample mean is zero  $t = \bar{x}$  is the requested root and the first proposition is proved.

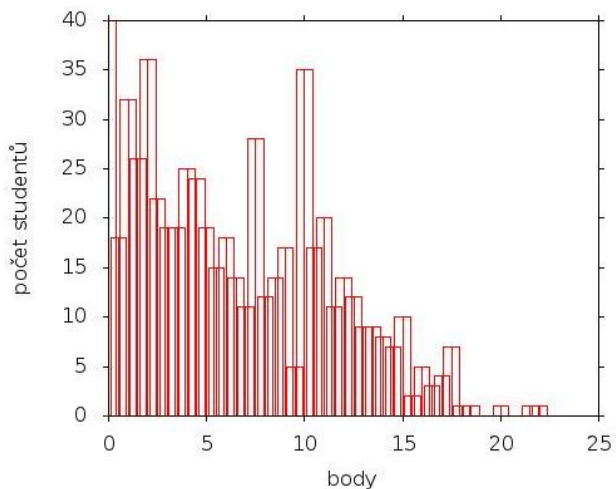
As for the second proposition, return to the definition of the median. For this purpose, rearrange the sum so that the first and the last summand is added, then the second and the last-but-one summand, etc. In the first case, this leads to the





We can notice that the mode of the values is 10, which, accidentally, was also the number of points necessary to pass the course. The mean of the obtained points is 9.48.

**10.A.8.** Here, we present column diagrams of the amounts of points of MB101 students in autumn 2010 (the very first semester of their studies). The first one corresponds to all students of the course; the second one does to those who (3 years later) successfully finished their studies and got the bachelor's degree.



Again, the results can be depicted in an alternative way:

expression  $|x_{(1)} - t| + |x_{(n)} - t|$ , and this is equal to the distance  $x_{(n)} - x_{(1)}$  provided  $t$  lies inside the range, and it is even greater otherwise. Similarly, the other pair in the sum gives  $x_{(n-1)} - x_{(2)}$  if  $x_{(2)} \leq t \leq x_{(n-1)}$ , and it is greater otherwise. Therefore, the minimality assumption leads to  $t = \tilde{x}$ .  $\square$

In practice, it is required to compare the variability of data sets of different statistical populations. For this purpose, it is convenient to relativize the scale, and so use the *coefficient of variation* of a data set  $x$ :

$$V_x = \frac{\sqrt{s_x^2}}{|\bar{x}|}.$$

This relative measure of variability can be perceived in percentage of the deviation with respect to the sample mean  $\bar{x}$ .

**10.1.7. Skewness of a data set.** If the values of a data set are distributed symmetrically around the mean value, then

$$\bar{x} = \tilde{x}$$

However, there are distributions where

$$\bar{x} > \tilde{x}.$$

This is common, for instance, with the distribution of salaries in a population where the mean is driven up by a few very large incomes, while much of the population is below the average.

A useful characteristic concerning this is the *Pearson coefficient*, given by

$$\beta = 3 \frac{\bar{x} - \tilde{x}}{s_x}.$$

It estimates the relative measure (the absolute value of  $\beta$ ) and the direction of the skewness (the sign). In particular, note that the standard deviation is always positive, so it is already the sign of  $\bar{x} - \tilde{x}$  which shows the direction of the skewness.

#### QUANTILE COEFFICIENTS OF SKEWNESS

More detailed information can be obtained from the *quantile coefficients of skewness*

$$\beta_p = \frac{x_{1-p} + x_p - 2\tilde{x}}{x_{1-p} - x_p},$$

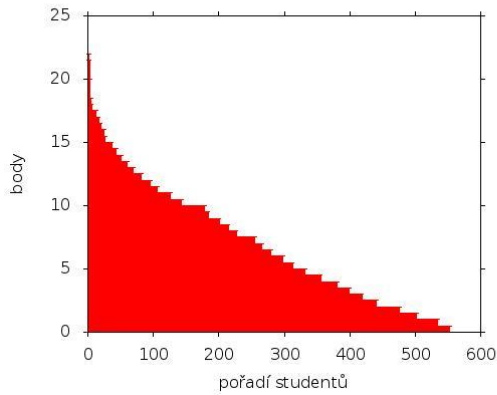
for each  $0 < p < 1/2$ . Their meaning is clear when the numerator is expressed as  $(x_{1-p} - \tilde{x}) - (\tilde{x} - x_p)$ .

In particular, the *quartile coefficient of skewness* is obtained when selecting  $p = 0.25$ .

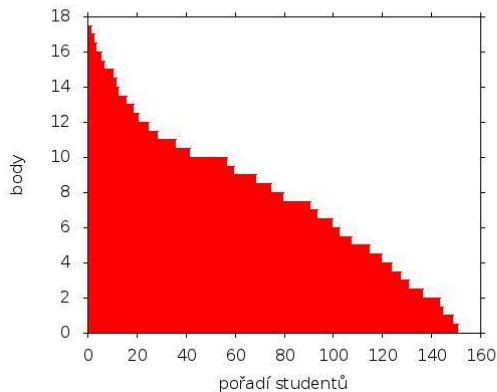
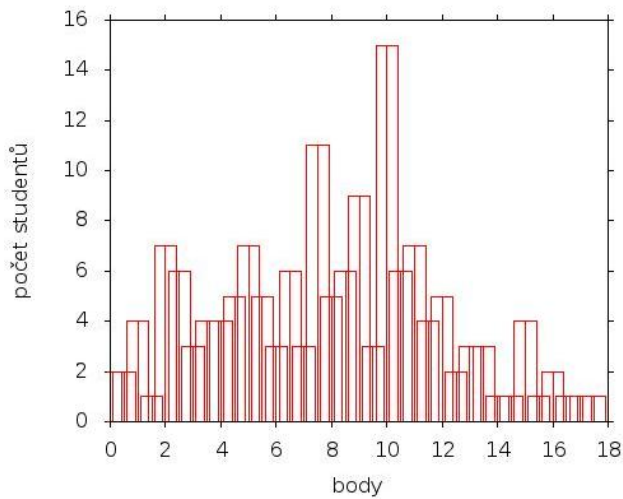
**10.1.8. Diagrams.** People's eyes are well suited for perceiving information with a complicated structure. That is why there exist many standardized tools for displaying statistical data or their correlations. One of



them is the *box diagram*.



And these are the graphs of amounts of points obtained by those students who continued their studies:

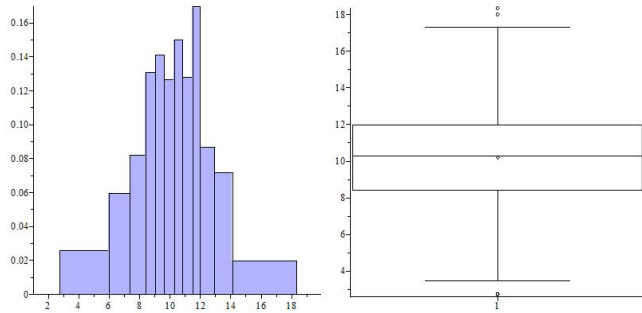


We can see that in the former case, the mode is equal to 0, while in the latter case, it is 10 again. The frequency distribution is close to the one of the MB104 course, which is recommended for the fourth semester.

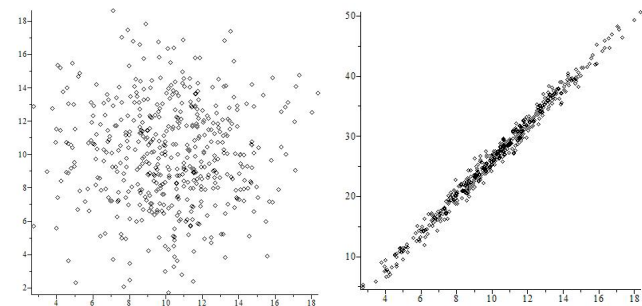
BOX DIAGRAM

The diagram illustrates a histogram and a box diagram of the same data set (normal distribution with mean equal to 10 and variance equal to 3,  $n = 500$ ).

The middle line is the median; the edges of the box are the quartiles; the “paws” show 1.5 of the interquartile range, but not more than the edges of the sample range. Potential outliers are indicated too.



Common displaying tools allow us to view potential dependencies of two data sets. For instance, in the left-hand diagram below, the coordinates are chosen as the values of two independent normal distributions with mean equal to 10 and variance equal to 3. In the right-hand illustration, the first coordinate is from the same data set, and the second coordinate is given by the formula  $y = 3x + 4$ . It is also perturbed with a small error.



**10.1.9. Covariance matrix.** Actually, the dependencies between several data sets associated to the same statistical units are at the core of our interest in many real world problems. When defining the variance in 10.1.6 above, we employed the euclidean distance, i.e. we evaluated the scalar product of the values of the square of distances from the mean with itself. Thus, having two vectors of data sets, we may define



**10.A.9.** A car was traveling from Brno to Prague at 160 km/h, and then back from Prague to Brno at 120 km/h. What was its average speed?

**Solution.** This is an example where one might think of using the arithmetic mean, which is incorrect. The arithmetic mean would be the correct result if the car spent the same period of time going at each speed. However, in this case, it traveled the same distance, not time, at each speed. Denoting by  $d$  the distance of Brno and Prague and by  $v_p$  the average speed, we obtain

$$\frac{d}{160} + \frac{d}{120} = \frac{2d}{v_p},$$

whence

$$v_p = \frac{2}{\frac{1}{160} + \frac{1}{120}} \doteq 137.14.$$

Therefore, the average speed is the harmonic mean (see 10.1.3) of the two speeds.

□

### B. Visualization of multidimensional data

The above examples were devoted to displaying one numerical characteristic measured for more objects (number of points obtained by individual students, for example). Graphical visualization of data helps us understand them better. However, how to depict the data if we measure  $p$  different characteristics,  $p \geq 3$ , of  $n$  objects. Such measurements cannot be displayed using graphs we have met.

**10.B.1.** One of the possible methods is the so-called *principal component analysis*. In this method, we use eigenvectors and eigenvalues (see 2.4.2) of the sample covariance matrix (see 10.2.35). We will use the following notation:

- random vectors of the measurement  
 $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T, i = 1, \dots, n,$
- the mean of the  $j$ -th component  
 $m_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, j = 1, \dots, p,$
- the sample variance of the  $j$ -th component  
 $s_j = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - m_j)^2, j = 1, \dots, p,$
- the vector of means  $\mathbf{m} = (m_1, \dots, m_p),$
- the sample covariance matrix  
 $\frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$   
 (note that each summand is a  $p$ -by- $p$  matrix).

The covariance matrix is symmetric, hence all its eigenvalues are real and its eigenvectors are pairwise orthogonal. Moreover, considering the eigenvectors of unit length, we can see that the eigenvalue corresponding to an eigenvector of the covariance matrix yields the variance of (the size of) the projection of the data onto this direction (the projection takes place in the  $p$ -dimensional space). The goal of this method is to find the direction (in the  $p$ -dimensional space of the measured characteristics) for which the variance of the projections is as great as possible. Thus, this direction corresponds to the eigenvector of the covariance matrix whose eigenvalue is the greatest one. The linear combination given

### COVARIANCE AND COVARIANCE MATRIX

Consider two data sets  $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n),$  and their means  $\bar{x}, \bar{y}.$  We define their *covariance* by the formula

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

If there are  $k$  sample sets  $x^{(1)} = (x_1^{(1)}, \dots, x_n^{(1)}), \dots, x^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)}),$  then their *covariance matrix* is the symmetric matrix  $C = (c_{ij})$  with  $c_{ij} = \text{cov}(x^{(i)}, x^{(j)}).$

Again, the *sample covariance* and *sample covariance matrix* are defined by the same formulae with  $n$  replaced by  $(n - 1).$

Clearly the covariance matrix has got the variances of the individual data sets on its diagonal.

In order to imagine what the covariance should say, consider the two possible behaviours of two data sets: (a) they will deviate from their means in a very similar way (comparing individually  $x_i$  and  $y_i$ ), (b) they behave very independently. In first case, we should assume that the signs of the deviations will mostly coincide and thus the sum in the definition will lead to a quite big positive number. In the other case the signs should be rather independent and thus the positive and negative contributions should effectively cancel each other in the covariance sum.

Thus we expect the data sets expressing independent features to be close to zero while the covariance of dependent sets should be far from zero. The sign of the covariance shows the character of the dependence. For example, the two sets of data depicted in the left hand diagram above had covariance about -0.11, while the covariance of the data from the right hand picture was about 25.9.

Similarly to the variance, we are often interested in normalized values. The *correlation coefficient* takes the covariance and divides it by the standard deviation of each of the data sets. In our two latter cases, the correlation coefficients are about -0.01 and 0.99. As expected, they very clearly indicate which of the data are correlated.

**10.1.10. Principal components analysis.** If we deal with statistics involving many parameters and we need to decide quickly about their similarity (correlation) with some given patterns, we might use a simple idea from linear algebra.

Assume we have got  $k$  data sets  $x^{(i)}.$  Since their covariance matrix  $C$  is symmetric, there is an orthonormal basis  $\underline{e}$  in  $\mathbb{R}^k$  such that in this basis the corresponding quadratic form given by  $C$  will enjoy a diagonal matrix. The relevant basis  $\underline{e}$  consists of the real eigenvectors  $e_i \in \mathbb{R}^k$  for the eigenvalues  $\lambda_i.$  The bigger is the absolute value  $|\lambda_i|,$  the bigger is the variation of the orthogonal projection  $\hat{x}$  of all the  $k$  data sets into this one-dimensional subspace spanned by  $e_i.$

Thus we may restrict ourselves to just this one data set  $\hat{x}$  and consider the statistics concerning this one set as representing the multi-parametric data sets  $x^{(i)}.$  Similarly we may

by the components of this vector is called the first principal component. The size of the projection onto this direction estimates the data quite well (the principal component can be viewed as a characteristic which substitutes for the  $p$  characteristics, i. e., it is a random vector with  $n$  components). If we subtract this projection from the data and consider the direction of the greatest variance again, we get the second principal component. Repeating this procedure further, we obtain the other principal components. The directions of the principal components correspond to the eigenvectors of the covariance matrix in decreasing order with respect to the size of the corresponding eigenvalues.

**10.B.2.** Find the first principal component of the following simple data and the vector which substitutes them: Five people were taken their height, little finger length, and index finger length. The measured data are shown in the following table (in centimeters).

**Solution.**

	Martin	Michael	Matthew	John	Peggy
index f.	9	11	8	8	8
little f.	7.5	8	6.3	6	6.5
height	186	187	173	174	167

The vectors of the collected data are:  $\mathbf{x}_1 = (9; 7.5; 186)$ ,  $\mathbf{x}_2 = (11; 8; 187)$ ,  $\mathbf{x}_3 = (8; 6; 173)$ ,  $\mathbf{x}_4 = (8; 6; 174)$ ,  $\mathbf{x}_5 = (8; 6.5; 167)$ . The covariance matrices of these vectors are:

$$\begin{pmatrix} 0.04 & 0.14 & 1.72 \\ 0.14 & 0.49 & 6.02 \\ 1.72 & 6.02 & 73.96 \end{pmatrix}, \quad \begin{pmatrix} 4.840 & 2.64 & 21.12 \\ 2.64 & 1.44 & 11.52 \\ 21.12 & 11.52 & 92.16 \end{pmatrix},$$

$$\begin{pmatrix} 0.641 & 0.640 & 3.521 \\ 0.640 & 0.640 & 3.52 \\ 3.521 & 3.52 & 19.36 \end{pmatrix}, \quad \begin{pmatrix} 0.641 & 0.640 & 2.721 \\ 0.640 & 0.640 & 2.72 \\ 2.721 & 2.72 & 11.56 \end{pmatrix},$$

$$\begin{pmatrix} 0.641 & 0.240 & 8.321 \\ 0.240 & 0.09 & 3.12 \\ 8.32 & 3.12 & 108.16 \end{pmatrix}.$$

The sample covariance matrix is then a quarter of their sum, i. e.,

$$S = \begin{pmatrix} 1.70 & 1.075 & 9.35 \\ 1.075 & 0.825 & 6.725 \\ 9.35 & 6.725 & 76.30 \end{pmatrix}$$

The eigenvalues of  $S$  are approximately 2.7, 312.2, and 0.38. The unit eigenvector corresponding to the greatest one is approximately  $(0.122; 0.09; 0.989)$ . Thus, the first principal component is  $(185.5; 186.8; 172.4; 173.4; 166.5)$ , which is not far from the people's heights.  $\square$

**10.B.3.** The students of a class had the following marks in various subjects:

also use several biggest eigen-values instead of one and reduce the dimension of our parameter space in this way. Finally, considering the unit length eigenvector  $(\alpha_1, \dots, \alpha_k)$  corresponding to the chosen eigenvalue  $\lambda$ , then the values  $\alpha_j$  provide the right coefficients in the orthogonal projection  $(x^{(1)}, \dots, x^{(k)}) \mapsto \hat{x} = \alpha_1 x^{(1)} + \dots + \alpha_k x^{(k)}$ .

See the exercise 10.B.2 for an illustration, together with another description how to proceed with the data in 10.B.1.

The latter approach is called the *principal component analysis*.

**10.1.11. Entropy.** We also need to describe the variability of data sets even with nominal types, for instance in statistical physics or information theory. The only thing at disposal is the class frequencies, so the principle of classical probability can be used (see the fourth part of chapter one). There, the relative frequency of the  $i$ -th class,  $p_i = \frac{n_i}{n}$ , is understood to be the probability that a random object belongs to this class.

The variance of ratio-type values with class frequencies  $n_j$  was given by the formula (see 10.1.6)

$$s_x^2 = \sum_{j=1}^m \frac{n_j}{n} (a_j - \bar{x})^2 = \sum_{j=1}^m p_j (a_j - \bar{x})^2,$$

where  $p_j$  denotes the (classical) probability that the value is in the  $j$ -th class. Therefore, it is a weighted mean of the adjusted values where the weight of the term  $(a_j - \bar{x})^2$  is  $p_j$ .

The variability of nominal values are expressed similarly (denote it by  $H_X$ ). Even though there are no numerical values  $a_j$  for the indices  $j$ , we can be interested in functions  $F$  that depend on the relative frequencies  $p_j$ . For a data set  $x$  we can define

$$H_X = \sum_{i=1}^n p_i F(p_i),$$

where  $F$  is an unknown function with some reasonable properties.

If the data set has only one value, i.e.  $p_k = 1$  for some  $k$  and otherwise  $p_j = 0$ , then we agree that the variability is zero, and so  $F(1) = 0$ .

Moreover,  $H_X$  is required to have the following property: If a data set  $Z$  consists of pairs of values from data sets  $X$  and  $Y$  (for example, one can observe eye colour and hair colour of people – statistical units), it is reasonable that the variability of  $Z$  be the sum of the variabilities, that is,  $H_Z = H_X + H_Y$ .

The relative class frequencies  $p_i$  for the values of the data set  $X$  and  $q_j$  for those of  $Y$  are known. The relative class frequencies for  $Z$  are then

$$r_{ij} = \frac{n_i m_j}{nm} = p_i q_j,$$

so we demand the equality (the ranges of the sums are clear from the context)

$$\sum_{i,j} p_i q_j F(p_i q_j) = \sum_i p_i F(p_i) + \sum_j q_j F(q_j).$$

Student id	Maths	Physics	History	English	PE
1	1	1	2	2	1
2	1	3	1	1	1
3	2	1	1	1	1
4	2	2	2	2	1
5	1	1	3	2	1
6	2	1	2	1	2
7	3	3	2	2	1
8	3	2	1	1	1
9	4	3	2	3	1
10	2	3	1	2	1

Find the first principal component of the following simple data and the vector which substitutes them.

**Solution.** The vectors of observation are  $\mathbf{x}_1 = (1, 1, 2, 2, 1), \dots, \mathbf{x}_{10} = (2, 3, 1, 2, 1)$ . The corresponding covariance matrices are:

$$\begin{pmatrix} 1.21 & 1.10 & -0.330 & -0.330 & 0.110 \\ 1.10 & 1. & -0.300 & -0.300 & 0.100 \\ -0.330 & -0.300 & 0.0900 & 0.0900 & -0.0300 \\ -0.330 & -0.300 & 0.0900 & 0.0900 & -0.0300 \\ 0.110 & 0.100 & -0.0300 & -0.0300 & 0.0100 \end{pmatrix}, \dots, \begin{pmatrix} 0.0100 & -0.100 & 0.0701 & -0.0300 & 0.0100 \\ -0.100 & 1. & -0.700 & 0.300 & -0.100 \\ 0.0701 & -0.700 & 0.490 & -0.210 & 0.0701 \\ -0.0300 & 0.300 & -0.210 & 0.0900 & -0.0300 \\ 0.0100 & -0.100 & 0.0701 & -0.0300 & 0.0100 \end{pmatrix}$$

The sample covariance matrix is

$$\begin{pmatrix} 0.99 & 0.44 & -0.078 & 0.26 & -0.01 \\ 0.44 & 0.89 & -0.22 & 0.22 & -0.11 \\ -0.078 & -0.22 & 0.45 & 0.23 & 0.03 \\ 0.26 & 0.22 & 0.23 & 0.45 & -0.078 \\ -0.01 & -0.11 & 0.033 & -0.0778 & 0.100 \end{pmatrix}.$$

Its dominant eigenvalue is about 13.68, and the corresponding unit eigenvector is approximately (0.70; 0.65; -0.13; 0.28; -0.07). Therefore, the principal component is (1.58; 2.73; 2.13; 2.93; 1.45; 1.93; 4.28; 3.48; 5.26; 3.71)  $\square$

Another possible method of visualization of multidimensional data is the so-called cluster analysis, but we will not go into further details here.

### C. Classical and conditional probability

In the first chapter, we met the so-called classical probability, see 1.4.1. Just to recall it, let us try to solve the following (a bit more complicated) problem:

**10.C.1.** Aleš wants to buy a new bike, which costs 5100 crowns. He has 2500 crowns left from organizing a camp. Aleš is no dope: he took 50 more crowns from his pocket money and went to the casino to play the roulette. Aleš always bets on red. This means that the probability of winning is 18/37 and the amount he wins is equal to the amount he has

Since  $p_i$  and  $q_j$  are relative frequencies, they sum to 1. So the right-hand side of the equality can be written as

$$\left(\sum_j q_j\right)\left(\sum_i p_i F(p_i)\right) + \left(\sum_i p_i\right)\left(\sum_j q_j F(q_j)\right),$$

leading to

$$\sum_{i,j} p_i q_j F(p_i q_j) = \sum_{i,j} p_i q_j (F(p_i) + F(q_j)).$$

This is satisfied by any constant multiple of a logarithm of any fixed base  $a > 1$ . It can be shown that no other continuous solution  $F$  exists.

Since  $p_i \leq 1$ ,  $\ln p_i \leq 0$ . The variability must be non-negative, so  $F$  is chosen to be a logarithmic function multiplied by  $-1$ . Such a choice also satisfies  $F(1) = 0$ , as desired.

### ENTROPY

The measure of variability of nominal values is expressed in terms of *entropy*. It is given by

$$H_X = -\sum_{i=1}^k \frac{n_i}{n} \ln\left(\frac{n_i}{n}\right),$$

where  $k$  is the number of sample classes. Sometimes (especially in information theory), the binary logarithm is used instead of the natural logarithm.

One often works with the quantity

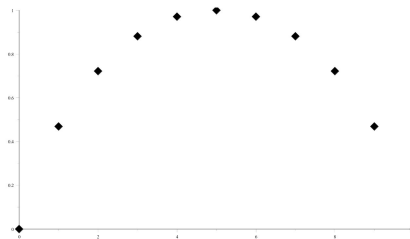
$$e^{H_X} = \prod_i p_i^{-p_i},$$

(or with another logarithm base).

In this form, for a data set  $X$  with  $k$  equal class frequencies, compute

$$e^{H_X} = \left(\left(\frac{1}{k}\right)^{-\frac{1}{k}}\right)^k = k,$$

which is independent of the sample size. The next illustration shows 2-based entropy  $y$  for the number of occurrences of letters  $a, b$  in 10-letter words consisting of these characters, and  $x$  is the number of occurrences of  $b$ .



Note that the maximum entropy 1 occurs for the same number of  $a$ 's and  $b$ 's, and indeed  $2^1 = 2$  as computed above.

The following illustration displays the entropy of 11 randomly chosen strings of length 10 made of 8 characters. The values are all much less than the theoretical maximal value of 3. This reflects the fact that the number of occurrences of the individual 8 characters cannot be equal (or it could happen with a very small probability if the length of the string was 8

bet. His betting strategy is as follows: The first time, he bets 10 crowns. Each time he has lost, he bets twice the previous bet (if he does not have enough money to make this bet, he leaves the casino, deeply depressed). Each time he has won, he bets 10 crowns again. What is the probability that, using this strategy, he wins the desired 2550 more crowns? (As soon as this happens, he immediately runs to buy the bike.)

**Solution.** First of all, we calculate how many times Aleš can lose in a row. If he bets 10 crowns the first time, then in order to bet  $n$  times, he needs

$$10+20+\dots+10\cdot 2^{n-1} = 10\cdot\left(\sum_{i=0}^{n-1} 2^i\right) = 10\cdot\left(\frac{2^n-1}{2-1}\right) = 10\cdot(2^n-1).$$

As we can see, the number 2550 is of the form  $10(2^n - 1)$ , for  $n = 8$ . This means that Aleš can bet eight times in a row regardless of the odds. He can never bet nine times in a row, because for that he would have to have  $10(2^9 - 1) = 5110$  crowns, which he will never reach (he stops betting as soon as he has 5100 crowns). Therefore, Aleš loses the whole game if and only if he loses eight consecutive bets. The probability of losing one bet is  $19/37$ ; hence, the probability of losing eight consecutive (independent) bets is  $(19/37)^8$ . Thus, the probability that he wins 10 crowns (using his strategy) is  $1 - (19/37)^8$ . In order to win 2550 crowns, he must win 255 times, and the probability of this is

$$\left(1 - \left(\frac{19}{37}\right)^8\right)^{255} \doteq 0.29.$$

Therefore, the probability of winning using his strategy is much lower than if he bet everything on red straightaway.  $\square$

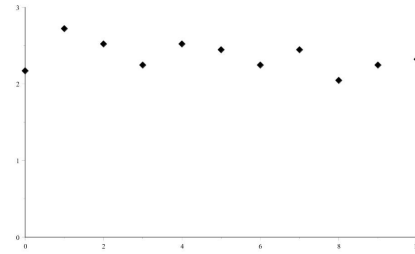
**10.C.2.** You could try to solve a slight modification of the above problem: Joe stops playing only if he loses all his money; if he still has some money, but not enough to bet twice the previous bet, he bets 10 dollars again.

We also met the conditional probability in the first chapter, see 1.4.8.

**10.C.3.** Let  $A, B$  be two events such that  $B$  is a disjoint union of events  $B_1, B_2, \dots, B_n$ . Using the definition of conditional probability (see 10.2.6), prove that

$$(1) \quad P(A|B) = \sum_{i=1}^n P(A|B_i)P(B_i|B)$$

or 16). If the same is done with, say, strings of length 10000, we would get very close to 3 (typically the difference would be in the order of  $10^{-3}$ , if the random string generator was good enough).



## 2. Probability

Before further reading, the reader is advised to go through the fourth part of chapter one (the subsection beginning on page 18). Back then, we worked mainly with classical finite probability. We defined the basics of a formalism which we extend now. The main extension is that the sample space  $\Omega$  can be infinite, even uncountable. Recall that when we talked about geometric probability at the end of the fourth part of chapter one, the sample space for description of an event was a part of the Euclidean space, and events were suitable subsets of it. All of those sets were uncountable.

Begin with a simple (infinite, yet still discrete) example, to which we return from time to time throughout this section.

**10.2.1. Why infinite sets of events?** Imagine an experiment where a coin is repeatedly tossed until it comes up heads. There are many questions to be asked about this experiment: What is the probability of tossing the coin at least 3 times? (or exactly 35 times, or at most 10 times, etc.)



The outcomes of this experiment can be considered in the form  $\omega_k \in \mathbb{N}_{\geq 1} \cup \{\infty\}$ , which could be read as “the coin comes up heads for the first time in the  $k$ -th toss”. Note that  $k = \infty$  is inserted, since the possibility that the coin always comes up tails must be allowed, too.

This problem is solved if the classical probability  $1/2$  of the coin coming up heads in one toss is used (and the same for tails). In the abstract model, the total number of tosses by any natural number  $N$  cannot be bounded. On the other hand, the probability that the coin always comes up tails in the first  $(k - 1)$  tosses out of the total number of  $n \geq k$  tosses is given by the fraction

$$\frac{2^{n-k}}{2^n} = 2^{-k},$$

where in the numerator, there is the number of favorable possibilities out of  $n$  independent tosses (i.e. the number of possibilities how to distribute two values into the  $n - k$  remaining positions), while in the denominator, there is the number of

**Solution.** First, note that the events  $A \cap B_1, A \cap B_2, \dots, A \cap B_n$  are also disjoint. Therefore, we can write

$$\begin{aligned} P(A|B_1 \cup \dots \cup B_n) &= \frac{P(A \cap (B_1 \cup \dots \cup B_n))}{P(B_1 \cup \dots \cup B_n)} = \\ &= \frac{P((A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n))}{P(B)} = \\ &= \frac{\sum_{i=1}^n P(A \cap B_i)}{P(B)} \cdot \frac{P(B_i)}{P(B)} = \\ &= \sum_{i=1}^n P(A|B_i)P(B_i|B). \end{aligned}$$

□

**10.C.4.** We have four bags with balls: In the first bag, there are four white balls. In the second bag, there are three white balls and one black ball. In the third bag, there are two white and two black balls. Finally, in the fourth bag, there are four black balls. We randomly pick a bag and take two balls out of it (without putting the first one back). Find the probability that

- the balls are of different colors;
- the second ball is white provided the first ball was white.

**Solution.** Since there is the same number of balls in each of the bags, any ball has the same probability of being taken (similarly for any pair of balls lying in the same bag). Therefore, we can solve this problem using classical probability

- Altogether, there are 24 pairs of balls that can be taken. Out of them, 7 consist of balls of different colors. Therefore, the wanted probability is  $7/24$ .
- Let  $A$  denote the event that the first ball is white and  $B$  denote the event that the second ball is white. Then,  $P(B \cap A)$  is the probability that both balls are white, and this is equal to  $10/24 = 5/12$  since there are 10 such pairs. Again, we can use classical probability to calculate  $P(A)$ : there are 16 balls in total, and 9 of them are white. Altogether, we have

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{5}{12}}{\frac{9}{16}} = \frac{20}{27}.$$

**Another solution.** The event  $A$  can be viewed as the union of three mutually exclusive events  $A_1, A_2, A_3$  that we took a white ball from the first, second, and third bag, respectively. Since there is the same number of balls in each of the bags, the probability of taking any (white) ball is also the same (independent of which ball it is), so we get  $P(A) = \frac{9}{16}$  and  $P(A_1|A) = \frac{\frac{4}{16}}{\frac{9}{16}} = \frac{4}{9}$ ,  $P(A_2|A) = \frac{3}{9} = \frac{1}{3}$ ,  $P(A_3|A) = \frac{2}{9}$ . Applying (5), we obtain

$$\begin{aligned} P(B|A) &= P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) = \\ &= P(B|A_1) \cdot \frac{P(A_1)}{P(A)} + P(B|A_2) \cdot \frac{P(A_2)}{P(A)} + P(B|A_3) \cdot \frac{P(A_3)}{P(A)} = \frac{\Omega \setminus \cup_{i \in I} A_i^c = \cap_{i \in I} A_i}{P(A)} = \\ &= 1 \cdot \frac{4}{9} + \frac{2}{3} \cdot \frac{3}{9} + \frac{1}{3} \cdot \frac{2}{9} = \frac{20}{27}. \end{aligned}$$

□

all possible outcomes. As expected, this probability is independent of the chosen  $n$ , and there is the  $\sum_{k=1}^{\infty} 2^{-k} = 1$ . Therefore, the probability of tossing only tails is zero.

Thus we can define probability on the sample space  $\Omega$  with sample points (outcomes)  $\omega_k$ , whose probability is  $2^{-k}$ . This leads to a probability according to the definitions below.

We return to this example throughout this section.

**10.2.2.  $\sigma$ -fields.** Work with a fixed non-empty set  $\Omega$ , which contains the possible outcomes of the experiment and which is called the *sample space*. The possible outcomes  $\omega \in \Omega$  are also called *sample points*. In probability models, not all subsets of outcomes need be admitted. In particular, the singletons  $\{\omega\}$  need not be considered. Those subsets whose probability we want to measure are required to satisfy the axioms of the so called  *$\sigma$ -algebras*.

The axioms listed below are chosen from a larger collection of natural requirements in a minimal form. The first one is based on the assumption that the universal event should be a measurable set. The second one is forced by the assumption that events can be negated. The third one reflects the necessity to examine the event of the occurrence of at least one event from a countably infinite collection. (For instance, in the example from the previous subsection, the coin is tossed only finitely many times, but there is no upper bound on the number of tosses.)

$\sigma$ -ALGEBRAS OF SUBSETS

A collection  $\mathcal{A}$  of subsets of the sample space is called a  *$\sigma$ -algebra* or  *$\sigma$ -field* and its elements are called *events* or *measurable sets* if and only if

- $\Omega \in \mathcal{A}$ , i.e., the sample space is an event;
- if  $A, B \in \mathcal{A}$ , then  $A \setminus B \in \mathcal{A}$ , i.e., the set difference of two events is also an event;
- if  $A_i \in \mathcal{A}$ ,  $i \in I$ , is a countable collection of events, then their union is also an event, i.e.,  $\cup_{i \in I} A_i \in \mathcal{A}$ .

As usual, the basic axioms imply simple corollaries which describe further (intuitively required) properties in the form of mathematical theorems. The reader should check carefully that both following properties hold.

- The complement  $A^c = \Omega \setminus A$  of an event  $A$  is again an event.
- The intersection of two events is again an event since for any two subsets  $A, B \subset \Omega$ ,

$$A \setminus (\Omega \setminus B) = A \cap B.$$

Actually, for any countable system of events  $A_i$ ,  $i \in I$ , the event  $\Omega \setminus \cup_{i \in I} A_i^c = \cap_{i \in I} A_i$  is also in the  $\sigma$ -algebra  $\mathcal{A}$ .

Altogether, a  $\sigma$ -algebra is a collection of subsets of the sample space which is closed with respect to set differences, countable unions, and countable intersections.

**10.C.5.** We have four bags with balls: In the first bag, there are four white balls. In the second bag, there are three white balls and one black ball. In the third bag, there are two white and two black balls. Finally, in the fourth bag, there are one white and three black balls. We randomly pick a bag and take a ball out of it, finding out that it is black. Then we throw away this bag, pick another one and take a ball out of it. What is the probability that it is white?

**Solution.** Similarly as in the above exercise, let  $A$  denote the event that the very first ball is black. This event can be viewed as the union of mutually exclusive events  $A_i$ ,  $i = 2, 3, 4$ , where  $A_i$  is the event of picking the  $i$ -th bag and taking a black ball from there. Again, the probability of picking any (black) ball is the same. Hence,  $P(A_2|A) = \frac{1}{6}$ ,  $P(A_3|A) = \frac{2}{6} = \frac{1}{3}$ , and  $P(A_4|A) = \frac{3}{6} = \frac{1}{2}$ . Let  $B$  denote the event that the second ball is white. If the thrown bag is the second one, then there are a total of 7 white balls remaining, so the probability of taking one of them is  $P(B|A_2) = \frac{7}{12}$  (we can use classical probability again because each of the bags contains the same number of balls, so any ball has the same probability of being taken). Similarly,  $P(B|A_3) = \frac{8}{12}$  and  $P(B|A_4) = \frac{9}{12}$ . Applying (5), we get that the wanted probability is

$$\begin{aligned} P(B|A) &= P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) + P(B|A_4)P(A_4|A) \\ &= \frac{7}{12} \cdot \frac{1}{6} + \frac{8}{12} \cdot \frac{1}{3} + \frac{9}{12} \cdot \frac{1}{2} = \frac{25}{36}. \quad \square \end{aligned}$$

**10.C.6.** We have four bags with balls: In the first bag, there are a white ball and a black ball. In the second bag, there are three white balls and one black ball. In the third bag, there are one white and two black balls. Finally, in the fourth bag, there are one white and three black balls. We randomly pick a bag and take a ball out of it, finding out that it is white. Then we throw away this bag, pick another one and take a ball out of it. What is the probability that it is white?

**Solution.** Similarly as in the above exercise, we view the event  $A$  of the first ball being white as the union of four mutually exclusive events  $A_1, A_2, A_3$ , and  $A_4$  that we take a white ball from the first, second, third, and fourth bag, respectively. The probability of taking a white ball out of the first bag is  $P(A_1) = \frac{1}{4} \cdot \frac{1}{2}$  (the probability of  $A_1$  is the product of the probability that we pick the first bag and the probability that we take a white ball from there); similarly,  $P(A_2) = \frac{1}{4} \cdot \frac{3}{4}$ ,  $P(A_3) = \frac{1}{4} \cdot \frac{1}{3}$ ,  $P(A_4) = \frac{1}{4} \cdot \frac{1}{4}$ .  $P(A) = P(A_1) + P(A_2) + P(A_3) + P(A_4) = \frac{11}{24}$ . Note that the probability  $P(A)$  cannot be calculated classically, i. e., by simply dividing the number of white balls by the total number of the balls, because, for instance, the probability of taking a white ball from the first bag is twice greater than from the fourth bag. As for the conditional probabilities, we have  $P(A_1|A) = P(A_1)/P(A) = \frac{3}{11}$ ,  $P(A_2|A) = \frac{9}{22}$ ,  $P(A_3|A) = \frac{2}{11}$ ,  $P(A_4|A) = \frac{3}{22}$ . Now, let  $B$  denote the event that we take another white ball after we have thrown away the first bag. We want to apply (5) again. It remains to compute  $P(B|A_i)$ ,  $i = 1, \dots, 4$ . The probability  $P(B|A_1)$  can be

**10.2.3. Probability space.** Now introduce probability in the mathematical model, recalling the concepts used already in the first chapter.

ELEMENTARY CONCEPTS

Use the following terminology in connection with events:

- the entire sample space  $\Omega$  is called the *universal event*; the empty set  $\emptyset \in \mathcal{A}$  is called the *null event*;
- the singletons  $\omega \in \Omega$  are called *elementary events* (note that  $\{\omega\}$  may not even be an event in  $\mathcal{A}$ );
- the intersection of events  $\cap_{i \in I} A_i$  corresponds to the *simultaneous occurrence* of all the events  $A_i, i \in I$ ;
- the union of events  $\cup_{i \in I} A_i$  corresponds to the *occurrence of at least one* of the events  $A_i, i \in I$ ;
- if  $A \cap B = \emptyset$ , then  $A, B \in \mathcal{A}$  are called *exclusive events* or *disjoint events*,
- if  $A \subset B$ , then the event  $A$  *implies* the event  $B$ ;
- if  $A \in \mathcal{A}$ , then the event  $B = \Omega \setminus A$  is called the *complementary event to  $A$*  and denoted  $B = A^c$ .

We have seen an example of probability defined on an infinite sample space in 10.2.1 above. In general, probability is comprehended as follows:

PROBABILITY

**Definition.** A *probability space* is the  $\sigma$ -algebra  $\mathcal{A}$  of subsets of the sample space  $\Omega$  on which there is a scalar function  $P : \mathcal{A} \rightarrow \mathbb{R}$  with the following properties:

- $P$  is non-negative, i.e.,  $P(A) \geq 0$  for all events  $A$ ;
- $P$  is countably additive, i.e.,

$$P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i),$$

for every countable collection of mutually exclusive events;

- the probability of the universal event is 1.

The function  $P$  is called the *probability function* on  $(\Omega, \mathcal{A})$ .

Immediately from the definition, the complementary event satisfies

$$P(A^c) = 1 - P(A).$$

In chapter one, theorems on addition of probabilities were derived. Although dealing with finite sample spaces, the arguments remain the same now. In particular, the *inclusion and exclusion principle* says for any finite collection of  $k$  events  $A_i$  that

$$\begin{aligned} P(\cup_{i=1}^k A_i) &= \sum_{i=1}^k P(A_i) - \sum_{i=1}^{k-1} \sum_{j=i+1}^k P(A_i \cap A_j) \\ &+ \sum_{i=1}^{k-2} \sum_{j=i+1}^{k-1} \sum_{\ell=j+1}^k P(A_i \cap A_j \cap A_\ell) \\ &- \dots + \\ &+ (-1)^{k-1} P(A_1 \cap A_2 \cap \dots \cap A_k). \end{aligned}$$



computed as the sum of the probabilities of the mutually exclusive events  $B_2, B_3, B_4$  (given  $A_1$ ) that the second white ball comes from the second, third, fourth bag, respectively. Altogether, we have

$$P(B|A_1) = P(B_2|A_1) + P(B_3|A_1) + P(B_4|A_1) = \frac{1}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{4}$$

Similarly,

$$P(B|A_2) = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{4} = \frac{13}{36}$$

$$P(B|A_3) = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{2}$$

$$P(B|A_4) = \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot \frac{1}{3} = \frac{19}{36}$$

Altogether, we get

$$P(B|A) = P(B|A_1)P(A_1|A) + P(B|A_2)P(A_2|A) + P(B|A_3)P(A_3|A) + P(B|A_4)P(A_4|A) = \frac{4}{9} \cdot \frac{3}{11} + \frac{13}{36} \cdot \frac{9}{22} + \frac{1}{2} \cdot \frac{2}{11} + \frac{19}{36} \cdot \frac{3}{22} = \frac{19}{44}$$

□

**10.C.7.** Two shooters shoot at a target, each makes two shots. Their respective accuracies are 80 % and 60 %. We have found two hits in the target. What is the probability that they belong to the first shooter?

**Solution.** The probability of hitting the target is  $4/5$  for the first shooter, and  $3/5$  for the second one. Consider the events:

$A \dots$  there are two hits in the target, both of the first shooter,

$B \dots$  there are two hits in the target.

Our task is to find  $P(B|A)$ . We can divide the event  $B$  into six disjoint events according to which shot(s) of each shooter was/were successful. We enumerate the events in a table and, for each of them, we compute its probability. This is easy as each of the events is the intersection of four independent events (results of the four shots). A hits is denoted by 1, a miss by 0.

	Shooter 1	Shooter 2	probability		
$B_1$	0	1	0	1	$\frac{1}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \cdot \frac{3}{5}$
$B_2$	0	1	1	0	$\frac{24}{25^2}$
$B_3$	1	0	1	0	$\frac{24}{25^2}$
$B_4$	1	0	0	1	$\frac{24}{25^2}$
$B_5$	1	1	0	0	$\frac{64}{25^2}$
$B_6$	0	0	1	1	$\frac{9}{25^2}$

Adding up the probabilities of these disjoint events, we get:

$$P(B) = \sum_{i=1}^6 P(B_i) = 169/625.$$

The reader should look back at 1.4.5 and think about the details.

**10.2.4. Independent events.** The definition of *stochastically independent events* also remains unchanged. It reflects the intuition that the probability of the simultaneous occurrence of independent events is equal to the product of the particular probabilities.

STOCHASTIC INDEPENDENCE

Events  $A, B$  are said to be stochastically independent if and only if

$$P(A \cap B) = P(A)P(B).$$

Of course, the universal event and the null event are stochastically independent of any event.

Recall that replacing an event  $A_i$  with the complementary event  $A_i^c$  in a collection of stochastically independent events  $A_1, A_2, \dots$ , again results in a collection of stochastically independent events, and (see 1.4.7, page 23)

$$P(A_1 \cup \dots \cup A_k) = 1 - P(A_1^c \cap \dots \cap A_k^c) = 1 - (1 - P(A_1)) \dots (1 - P(A_k)).$$

Classical finite probability remains the fundamental example of probability, used as the inspiration during creation of the mathematical model. Recall that in this case,  $\Omega$  is a finite set, the  $\sigma$ -algebra  $\mathcal{A}$  is the collection of all subsets of  $\Omega$ , and the *classical probability* is the probability space  $(\Omega, \mathcal{A}, P)$  with probability function  $P : \mathcal{A} \rightarrow \mathbb{R}$ ,

$$P(A) = \frac{|A|}{|\Omega|}.$$

This corresponds precisely to the intuition about the relative frequency  $p_A$  of an event  $A$  when drawing a random element from the sample set  $\Omega$ .

This definition of probability guarantees reasonable behaviour of monotone sequences of events:

**10.2.5. Theorem.** Consider a probability space  $(\Omega, \mathcal{A}, P)$  and a non-decreasing sequence of events  $A_1 \subset A_2 \subset \dots$ . Then,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \rightarrow \infty} P(A_i).$$

Similarly, if  $A_1 \supset A_2 \supset A_3 \supset \dots$ , then

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \rightarrow \infty} P(A_i).$$

**PROOF.** The considered union  $A = \bigcup_{i=1}^{\infty} A_i$  can be rewritten in terms of mutually exclusive events

$$\tilde{A}_i = A_i \setminus A_{i-1},$$

defined for all  $i = 2, 3, \dots$ . Set  $\tilde{A}_1 = A_1$ . Then,

$$P(A) = P\left(\bigcup_{i=1}^{\infty} \tilde{A}_i\right) = \sum_{i=1}^{\infty} P(\tilde{A}_i) = \lim_{k \rightarrow \infty} \sum_{i=1}^k P(\tilde{A}_i).$$

Now, we can compute the conditional probability, using the formula of subsection 10.2.6:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B_5)}{P(B)} = \frac{\frac{64}{625}}{\frac{169}{625}} = \frac{64}{164} \cdot 0.38.$$

□

**10.C.8.** We toss a coin. If it comes up heads, we put a white ball into an (initially empty) bag; otherwise we put a black ball there. This is repeated  $n$  times. Then, we take a ball randomly from the bag (without replacement). Suppose it is white. What is the probability that another ball we take randomly from the bag is black?

**Solution.** We will solve the problem for a general (possibly biased) coin. In particular, we assume that the individual tosses are independent and that there exists a fixed probability of the coin coming up heads, which we denote  $p$ . The event “a ball in the bag is white” corresponds to the event “the coin came up heads in the corresponding toss”. Since the first ball was white, we deduce that  $p > 0$ . We can also see that the probability space “taking a random ball from the bag” is isomorphic to the probability space “tossing a coin”. Since we assume that the individual tosses are independent, we also get the independence of the colors of the selected balls. This leads to the conclusion that the probability in question is  $1-p$ .

Is this reasoning correct? Do we not expect the probability of taking a black ball to be greater than  $1-p$ ? See, there were approximately  $np$  white and  $n(1-p)$  black balls in the bag, so if we had removed one white ball, the probability of selecting a black one should increase, shouldn't it? Before reading further, try to figure out which (if any) of these two presented reasonings is correct, and whether the probability is also dependent on  $n$  (the number of balls in the bag before any were removed).

Now, we select a more sophisticated approach to the problem. Let  $B_i$  denote the event “there were  $i$  white balls in the bag” (before any were removed),  $i \in \{0, 1, 2, \dots, n\}$ . Further, let  $A$  denote the event “the first ball is white” and  $C$  denote the event “the second ball is black”. Actually, the event  $B_i$  says that the coin came up heads  $i$  times out of  $n$ ; hence, its probability is

$$P(B_i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

The conditional probability of taking a white ball provided there are exactly  $i$  white balls in the bag is equal to

$$P(A|B_i) = \frac{i}{n}.$$

We are interested in the probability of  $C$ , knowing that  $A$  has occurred, i. e., we want to know  $P(C|A)$ . Since the events  $B_i$  are pairwise disjoint, this is also true for the events  $C \cap B_i$ . Since  $C$  can be decomposed as the disjoint union  $\bigcup_{i=0}^n (C \cap$

For the finite sums,

$$\sum_{i=1}^k P(\tilde{A}_i) = P(A_1) + \sum_{i=2}^k (P(A_i) - P(A_{i-1})) = P(A_n)$$

by the assumptions  $A_{i-1} \subset A_i$ . This proves the first part of the theorem.

In the second part, consider the complements  $B_i = A_i^c$  instead of the events  $A_i$ . They satisfy the assumptions of the first part of this theorem. Then, the complement of the considered intersection is

$$B = A^c = \left( \bigcap_{i=1}^{\infty} A_i \right)^c = \bigcup_{i=1}^{\infty} B_i.$$

The desired statement follows from the fact that

$$P(A) = 1 - P(B) = 1 - \lim_{i \rightarrow \infty} P(B_i) = \lim_{i \rightarrow \infty} (1 - P(B_i))$$

which completes the proof. □

**10.2.6. Conditional probability.** Consider the following problem: On average, 40% of students succeed in course  $X$  and 80% of students succeed in course  $Y$ . If a random student is enrolled in both these courses saying that he has passed one of them (but we overhear which one), what is the probability that he has meant course  $X$ ?



As mentioned in subsection 1.4.8 (page 24), such problems can be formalized in the way described below. (We shall come back to the solution of the latter problem in 10.3.12.)

CONDITIONAL PROBABILITY

**Definition.** Let  $H$  be an event with non-zero probability in the  $\sigma$ -algebra  $\mathcal{A}$  of a probability space  $(\Omega, \mathcal{A}, P)$ . The *conditional probability*  $P(A|H)$  of an event  $A \in \mathcal{A}$  with respect to the hypothesis  $H$  is defined as

$$P(A|H) = \frac{P(A \cap H)}{P(H)}.$$

The definition corresponds to the intuition from the classical probability that the probability of events  $A$  and  $H$  occurring simultaneously, provided the event  $H$  has occurred, is  $P(A \cap H)/P(H)$ .

Directly from the definition, the hypothesis  $H$  and the event  $A$  are independent if and only if  $P(A) = P(A|H)$ .

At first sight, it may seem that introducing conditional probability does not add anything new. Actually, it is a very important type of approach which is needed in statistics as well. The hypothesis can be the a priori probability (i.e. the prior belief assumed beforehand), and the resulting probability is said to be posterior (i.e., it is considered to be a consequence of the assumption). This is the core of the Bayesian approach to statistics as is seen later.

The definition also implies the following result.

$B_i$ ), we can write

$$\begin{aligned} P(C|A) &= P\left(\bigcup_{i=0}^n (C \cap B_i) | A\right) = \sum_{i=0}^n \frac{P((C \cap B_i) \cap A)}{P(A)} = \\ &= \frac{1}{P(A)} \sum_{i=0}^n P(C \cap (A \cap B_i)) = \\ &= \frac{1}{P(A)} \sum_{i=0}^n P(A \cap B_i) P(C|A \cap B_i) = \\ &= \frac{1}{P(A)} \sum_{i=0}^n P(B_i) P(A|B_i) P(C|A \cap B_i). \end{aligned}$$

We use the law of total probability and substitute for  $P(A)$ , which leads to

$$\begin{aligned} P(C|A) &= \frac{\sum_{i=0}^n P(B_i) P(A|B_i) P(C|A \cap B_i)}{P(A)} = \\ (1) \quad &= \frac{\sum_{i=0}^n P(B_i) P(A|B_i) P(C|A \cap B_i)}{\sum_{i=0}^n P(B_i) P(A|B_i)}. \end{aligned}$$

This formula is sometimes called the *second Bayes' formula*; it holds in general provided the space  $\Omega$  is a disjoint union of the events  $B_i$ .

Since we tossed the coin at least once, we have  $n \geq 1$ . Now, we can calculate:

$$\begin{aligned} \sum_{i=0}^n P(B_i) P(A|B_i) &= \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} \cdot \frac{i}{n} = \\ &= \sum_{i=1}^n \frac{(n-1)!}{(i-1)!(n-i)!} p^i (1-p)^{n-i} = \\ &= \sum_{i=0}^{n-1} \frac{(n-1)!}{i!(n-i-1)!} p^{i+1} (1-p)^{n-i-1} = \\ &= p \sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-1-i} = \\ &= p(p + (1-p))^{n-1} = p, \\ \sum_{i=0}^n P(B_i) P(A|B_i) P(C|A \cap B_i) &= \\ &= \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} \cdot \frac{i}{n} \cdot \frac{n-i}{n-1} = \\ &= \sum_{i=1}^{n-1} \frac{(n-2)!}{(i-1)!(n-i-1)!} p^i (1-p)^{n-i} = \\ &= \sum_{i=0}^{n-2} \frac{(n-2)!}{i!(n-2-i)!} p^{i+1} (1-p)^{n-i-1} = \\ &= p(1-p) \sum_{i=0}^{n-2} \binom{n-2}{i} p^i (1-p)^{n-2-i} = \end{aligned}$$

**Lemma.** Let an event  $B$  be the union of mutually exclusive events  $B_1, B_2, \dots, B_n$ . Then,

$$(1) \quad P(A|B) = \sum_{i=1}^n P(A|B_i) P(B_i|B)$$

**PROOF.** The events  $A \cap B_1, A \cap B_2, \dots, A \cap B_n$  are also mutually exclusive. Therefore,

$$\begin{aligned} P(A|B_1 \cup \dots \cup B_n) &= \frac{P(A \cap (B_1 \cup \dots \cup B_n))}{P(B_1 \cup \dots \cup B_n)} = \\ &= \frac{P((A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n))}{P(B)} = \\ &= \sum_{i=1}^n \frac{P(A \cap B_i) P(B_i)}{P(B_i) P(B)} = \\ &= \sum_{i=1}^n P(A|B_i) P(B_i|B). \quad \square \end{aligned}$$

Consider the special case  $B = \Omega$ . Then, the events  $B_i$  can be considered the “possible states of the universe”,  $P(A|B_i)$  expresses the probability of  $A$  provided the universe is in its  $i$ -th state, and  $P(B_i|\Omega) = P(B_i)$  is the probability of the universe being in its  $i$ -th state. By the above lemma,

$$P(A) = P(A|\Omega) = \sum_{i=1}^n P(A|B_i) P(B_i).$$

This formula is called the *law of total probability*.

**10.2.7. Bayes' theorem.** Simple rearrangement of the conditional probability formula leads to

$$P(A \cap B) = P(B \cap A) = P(A)P(B|A) = P(B)P(A|B).$$

There are two important corollaries:

BAYES' RULES

**Theorem.** The probabilities of events  $A$  and  $B$  satisfy

$$\begin{aligned} (1) \quad P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \\ (2) \quad P(A|B) &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}. \end{aligned}$$

The first proposition is called the *inverse probability formula*. The second proposition is called the first *Bayes' formula*.

**PROOF.** The first statement is a mere rearrangement of the formula above the theorem. To obtain the second statement, note that

$$P(B) = P(B \cap A) + P(B \cap A^c).$$

Applying the law of total probability,  $P(B) = P(A)P(B|A) + P(A^c)P(B|A^c)$  can be substituted into the inverse probability formula, thereby obtaining the second statement of the theorem.  $\square$

these two subsections are now moved to Chapter one, modify/extend here ...

$$= \begin{cases} p(1-p), & n > 1 \\ 0, & n = 1. \end{cases}$$

Substituting this into the second Bayes' formula, we obtain the wanted probability

$$P(C|A) = \begin{cases} 0, & n = 1, \\ 1-p, & n > 1. \end{cases}$$

Thus, the simple reasoning about the probability spaces being isomorphic led to the correct result. The second reasoning was wrong because it omitted the fact that since the first ball was white, the expected number of white balls in the bag (before removing the first one) was greater than  $np$ . The calculation highlights the singular case  $n = 1$ .  $\square$

**10.C.9.** Once upon a time, there was a quiz where the first prize was Ferrari 599 GTB Fiorano. The contestant who won the final round was taken into a room where there were three identical doors. Behind two of them, there were goats, while the third one contained the car. In order to win the car, the contestant had to guess the correct door. First, the contestant pointed at one of the three doors. Then, an assistant opened one of the other two doors behind which there was a goat. Now, the contestant is given the option to change his guess. Should he do so?

**Solution.** Of course, we assume that the contestant wants to win the car. First of all, try to examine your intuition for random events. For example, you can reason as follows: "One of the two remaining doors contains the car, each with the same probability. Therefore, it does not matter which door we choose." Or: "The probability of choosing the correct door at the beginning is  $\frac{1}{3}$ . The shown goat changes nothing, so the probability that the guess is wrong is  $\frac{2}{3}$ . Therefore, we should change the door, thereby winning by  $\frac{2}{3}$ ."

Apparently, it is wise to change the door only if the probability of the car being behind that door is greater than behind the initially chosen one. We consider the following events:  $H$  stands for "the initial guess is correct",  $A$  stands for "we have changed the door", and  $C$  for "we have won". We are thus interested in the probabilities  $P(C|A)$  and  $P(C|A^c)$ .

First, we choose one of three doors, and the Ferrari is behind one of them, so

$$P(H) = \frac{1}{3}, \quad P(H^c) = 1 - \frac{1}{3} = \frac{2}{3}.$$

We assume that the event of changing the door is independent of the original guess, hence

$$P(A|H) = P(A|H^c) = P(A), \quad P(A^c|H) = P(A^c|H^c) = P(A^c).$$

If the original guess is correct and it is changed, then we surely lose; while if it is originally wrong and then it is changed, then we surely win. Therefore, we have

$$\begin{aligned} P(C|A \cap H) &= 0 = P(C|A^c \cap H^c), \\ P(C|A^c \cap H) &= 1 = P(C|A \cap H^c). \end{aligned}$$

Bayes' rule is sometimes formulated in a somewhat more general form, proved similarly as in (??):

Let the sample space  $\Omega$  be the union of mutually exclusive events  $A_1, \dots, A_n$ . Then, for any  $i \in \{1, \dots, n\}$ ,

$$(3) \quad P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^n P(B|A_k)P(A_k)}.$$

**10.2.8. Example and remarks.** Now, the introductory question from 10.2.6 can be dealt with easily. Consider the event  $A$  which corresponds to "the student having passed an exam" and the event



$B$  which corresponds to "the exam in question concerning course  $X$ ". Assume that the probabilities of the exam concerning either course are the same, i.e.,  $P(B) = P(B^c) = 0.5$ . While the wanted probability  $P(B|A)$  is unclear, the probability  $P(A|B) = 0.4$  is given, as well as  $P(A|B^c) = 0.8$ .

This is a typical application of Bayes' formula ??(??). There is no need to calculate  $P(A)$  at all:

$$\begin{aligned} P(B|A) &= \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)} = \\ &= \frac{0.5 \cdot 0.4}{0.5 \cdot 0.4 + 0.5 \cdot 0.8} = \frac{1}{3}. \end{aligned}$$

In order to better understand the role of the prior probability hypothesis, here is another example.

Consider a university using entrance exams with the following reliability: 99% of intelligent people pass them, while concerning non-intelligent people, only 0.5% are able to pass. It is desired to find the probability that a random student (accepted applicant) of the university is intelligent.

Thus, let  $A$  be the event "a random person is intelligent" and  $B$  be the event "the person passed the exams successfully". Using Bayes' formula, the probability that  $A$  occurs provided  $B$  has occurred can be computed. It is only necessary to supply the general probability  $p = P(A)$  that a random applicant is intelligent.

$$P(A|B) = \frac{p \cdot 0.99}{p \cdot 0.99 + (1-p) \cdot 0.005}.$$

The following table presents the result for various values of  $p$ . The first column corresponds to the case that every other applicant is intelligent, etc.

$p$	0.5	0.1	0.05	0.01	0.001	0.0001
$P(A B)$	0.99	0.96	0.91	0.67	0.17	0.02

Therefore, if every other applicant is intelligent, then 99% of the students are intelligent. If only 1% of the population meets an expectation of "intelligence" and the applicants form a good random sample, then only about two thirds of the students are intelligent, etc.

Consider similar tests for the occurrence of a disease, say HIV. There may be a test with the same reliability as the one above and use it to test all students that are present at the university. In this case, assume that the parameter  $p$  is close to the one for the entire population (say 1 out of 10000 people is infected, on average), which corresponds to the last column

It follows from the second Bayes' formula (1) that

$$\begin{aligned}
 P(C|A) &= \\
 &= \frac{P(H)P(A|H)P(C|A \cap H) + P(H^c)P(A|H^c)P(C|A \cap H^c)}{P(A)} \\
 &= P(H^c) = \frac{2}{3}
 \end{aligned}$$

and, analogously,

$$\begin{aligned}
 P(C|A^c) &= \\
 &= \frac{P(H)P(A^c|H)P(C|A^c \cap H) + P(H^c)P(A^c|H^c)P(C|A^c \cap H^c)}{P(A^c)} \\
 &= P(H) = \frac{1}{3}.
 \end{aligned}$$

We have thus obtained  $P(C|A) > P(C|A^c)$ , which means that it is wise to change the door.

Note that the solution is based upon the assumption that the assistant *deliberately* opens a door behind which there is a goat. If the contestant believes it was an accident or if instead, say, he happens to see (or hear) a goat behind one of the two not chosen doors, then the first reasoning is correct and the probability remains to be  $\frac{1}{2}$ .  $\square$

**10.C.10.** We have two bags. The first one contains two white and two black balls, while the second one contains one white and two black balls. We randomly select one of the bags and take two balls out of it (without replacement). What is the probability that the second ball is black provided the first one is white?  $\circ$

### D. What is probability?

First of all, recall the geometric probability, which was introduced in ??.

**10.D.1. Buffon's needle.** A plane is covered with parallel lines, creating bands of width  $l$ . Then, a needle of length  $l$  is thrown onto the plane. What is the probability that the needle crosses one of the lines?

**Solution.** The position of the needle is given by two independent parameters: the distance  $d$  of the needle's center from the closest line ( $d \in [0, l/2]$ ) and the angle  $\alpha$  ( $\alpha \in [0, \pi/2]$ ) between the lines and the needle's direction. The needle crosses one of the lines if and only if  $l/2 \sin \alpha > d$ . The space of all events  $(\alpha, d)$  is a rectangle  $\pi/2 \times l/2$ . The favorable events  $(\alpha, d)$  (i. e. those for which  $l/2 \sin \alpha > d$ ) correspond to those points in the rectangle which lie under the curve  $l/2 \sin \alpha$  ( $\alpha$  being the variable of the  $x$ -axis). By 6.2.20, the area of the figure is

$$\int_0^{\pi/2} \frac{l}{2} \sin \alpha \, d\alpha = \frac{l}{2}.$$

Thus, the wanted probability is (see ??)

$$\frac{\frac{l}{2}}{\frac{\pi}{2} \cdot \frac{l}{2}} = \frac{2}{\pi}.$$

of the table above. Clearly the result of the test is catastrophically unreliable. Only about 2% of the students who are tested positive are really infected!

Note that the problem with both tests is the same one. It is clear that real entrance exams require good selectivity and reliability. So the university marketing must ensure that the actual applicants do not provide a good random sample of population. Perhaps the university should try to discourage "non-intelligent" people from applying and thus secure a sufficiently low number of such applicants. With diseases, even the very rare occurrence of healthy people tested positively can be devastating. If the test is improved so that it is 100% reliable for positive people, it would have almost no impact on the resulting probabilities in the table.

Thus, if a person is tested positive when diagnosing a rare disease, it is necessary to make further tests. Then, the result  $P(A|B)$  of the first test plays the role of the prior probability  $P(A)$  during the second test, etc. This approach allows one to "cumulate the experience".

**10.2.9. Borel sets.** In practice, the probability of events which are expressed by questioning whether some numerical quantity falls into a given interval is of interested. We illustrate this on the example dealing with the results of students in a given course, measured for instance by the number of points in a written exam (cf. 10.1.1).

On one hand, there is only a finite number of students, and there are only a finite number of possible results (say, the numbers of points in the written exam can be the integers 0 through 20). On the other hand, imagining the results of the students as an analogy to independent rolls of a regular die is inappropriate. Even if a regular 21-hedron would exist (it cannot, see chapter 13); that would be somewhat weird.

Thus it is better to focus on the assessing function  $X : \Omega \rightarrow \mathbb{R}$  in the sample space  $\Omega$  of all students and model the probability that its value falls into a fixed interval when a random student is picked. For instance, if the table transferring points into marks A through F is fixed, the probability that the student obtained an A or a B can be modeled.

In the case of a reasonable course, we should expect that the most probable results are somewhere in the middle of the "interval of success", while the ideal result of the full number of points is not very probable. Similarly, if many values of  $X$  lie in the interval of failure, this may be at most universities perceived as a significant failure of the lecturer. This is a typical example of the random variables or random vectors, as defined below (it depends whether the result of just one or several students is chosen randomly).

One way to proceed is to model the behaviour of  $X$  as probability defined for all intervals. This requires the following  $\sigma$ -algebra.<sup>1</sup>

<sup>1</sup>In this connection, we also talk about the  $\sigma$ -algebra of *Borel-measurable sets* on  $\mathbb{R}^k$ , and then the following definition says that random variables are *Borel-measurable functions*.  $\square$

The following (known) problem, which also deals with geometric probability, illustrates that we must be cautious about what is assumed to be “clear”.

**10.D.2. Bertrand’s paradox.** What is the probability that a random chord of a given circle is longer than the side of an equilateral triangle inscribed into the circle?

**Solution.** We will show three ways how to find “this” probability.

1) Every chord is determined by its center. Thus, a random choice of the chord is given by a random choice of the center. The chord is greater than the side of the inscribed equilateral triangle if and only if its center lies inside the concentric circle with half radius. The center is chosen “randomly” from the whole inside of the circle. Therefore, the probability that it will lie in the inner disc is given by the ratio of the areas of these discs, which is  $\frac{1}{4}$ .

2) Unlike above, we claim that the wanted probability does not change if the direction of the chord is fixed. Then, the centers of such chords lie on a fixed diameter of the circle. The favorable centers are those which lie inside the inner circle (see 1)), i. e., inside a fixed diameter of the inner circle. The ratio of the diameters is 1 : 2, hence the wanted probability is  $\frac{1}{2}$ .

3) Now, we observe that a chord is determined by its endpoints (which must lie on the circle). Let us fix one of the endpoints (call it  $A$ )—thanks to the apparent symmetry, this should not affect the resulting probability. Then, the chord satisfies the given condition if and only if the other endpoint lies on the shorter arc  $BC$ , where  $ABC$  is the inscribed equilateral triangle. However, the length of this arc is one third of the length of the entire circle, which means that the wanted probability is equal to  $\frac{1}{3}$ .

How is it possible that we came to three different probabilities? It is caused by a hidden ambiguity in the statement of the problem. It is necessary to specify what exactly it means to choose a chord “randomly”. Each of the three results is correct provided the chord is chosen in the corresponding way. However, these ways are not equivalent; this is apparent not only from the different results, but also from the distribution of the chords’ centers. In the first case, they are distributed uniformly throughout the inside of the circle. In the second and third cases, the centers are concentrated more towards the center of the circle.

□

**10.D.3. Two envelopes.** There are two envelopes, each contains a certain amount of money. We know that the amount in one of them is twice as great as in the other one. We can choose either of the envelopes (and take its contents). As soon as we choose one, we are allowed to change our mind and take the other envelope instead. Is it advantageous to do so?

**Solution.** At the first sight, it must not matter which envelope we choose. The probability of choosing the one which contains more is 1/2, so it is no good to change our choice.

BOREL SETS

The *Borel sets* in  $\mathbb{R}$  are all those subsets that can be obtained from intervals using complements, countable unions, and countable intersections.

More generally, on the sample space  $\Omega = \mathbb{R}^k$ , one considers the smallest  $\sigma$ -algebra  $\mathcal{B}$  which contains all  $k$ -dimensional intervals.

The sets in  $\mathcal{B}$  are called the *Borel sets* on  $\mathbb{R}^k$ .

**10.2.10. Random variables.** The probabilities of the individual intervals in the Borel algebra are usually given as follows. Consider a numerical quantity  $X$  on any sample space, that is, a function  $X : \Omega \rightarrow \mathbb{R}$ . Since it is desired to work with the probability of  $X$  taking on values from any fixed interval, the probability space and the properties of the function  $X$  have to allow this.



Notice that working with finite probability spaces where all subsets are events, every function  $X : \Omega \rightarrow \mathbb{R}$  is a random variable in the following sense.

RANDOM VARIABLES AND VECTORS

**Definition.** A *random variable*  $X$  on a probability space  $(\Omega, \mathcal{A}, P)$  is a function  $X : \Omega \rightarrow \mathbb{R}$  such that the inverse image  $X^{-1}(B)$  lies in  $\mathcal{A}$  for every Borel set  $B \in \mathcal{B}$  on  $\mathbb{R}$ . The real-valued function  $P_X(B) = P(X^{-1}(B))$  defined on all intervals  $B \subset \mathbb{R}$  is called the (*probability*) *distribution of a random variable*  $X$ .

A *random vector*  $X = (X_1, \dots, X_k)$  on  $(\Omega, \mathcal{A}, P)$  is a  $k$ -tuple of random variables  $X_i : \Omega \rightarrow \mathbb{R}$  defined on the same probability space  $(\Omega, \mathcal{A}, P)$ .

If intervals  $I_1, \dots, I_k$  in  $\mathbb{R}$  are chosen, then the probability of simultaneous occurrence of all of the  $k$  events  $X_i \in I_i$  must exist. Thus, as in the scalar case, there is a real-valued function defined on the  $k$ -dimensional intervals  $B = I_1 \times \dots \times I_k$ ,  $P_X(B) = P(X^{-1}(B))$  (and thus also for all Borel sets  $B \subset \mathbb{R}^k$ ). It is called the *probability distribution of the random vector*  $X$ .

**10.2.11. Distribution function.** The distribution of random variables is usually given by a rule which shows how the probability grows as the interval  $B$  is extended.

In particular, consider the intervals  $I$  with endpoints  $a, b$ ,  $-\infty \leq a \leq b \leq \infty$ . Denote  $P(a < X < b)$  the probability of  $X$  lying in  $I = (a, b)$ , or  $P(X < b)$  if  $a = -\infty$ ; and analogously for other types of intervals. In the special case of a singleton, write  $P(X = a)$ .

In the case of a random vector  $X = (X_1, \dots, X_k)$ , write  $P(a_1 < X_1 < b_1, \dots, a_k < X_k < b_k)$  for the probability of simultaneous occurrence of the events where the values of  $X_i$  fall into the corresponding intervals (which may also be closed, unbounded, etc.).

However, consider the following reasoning: the envelope we have chosen contains  $a$ . Therefore, the other one contains  $a/2$  or  $2a$ , each with probability  $1/2$ . This means that if we change the envelope, then we get  $a/2$  with probability  $1/2$  and  $2a$  with probability  $1/2$ , i. e., the expected outcome is

$$\frac{1}{2} \frac{a}{2} + \frac{1}{2} 2a = \frac{5}{4} a.$$

Therefore, it is wise to change the envelope. What is wrong with this reasoning?

There are several issues. Mainly, it is not generally true that if there is amount  $a$  in one of the envelopes, then the second one contains  $a/2$  with probability  $1/2$ . This depends on the initial distribution of the amounts that have been put into the envelopes, which is not precisely stated in the problem.

However, the paradox is rooted not only in the concealed a priori distribution. There are (even discrete) distributions for which the choice of changing the envelope always produces greater expected outcome than that of not changing it. Nevertheless, any distribution with this property must have infinite expected value (if the expectation is finite, then there is always a value which, when seen in the envelope, it is more advantageous to keep), so it is dubious to say that it is better to get “greater” infinity on average.  $\square$

### E. Random variables, density, distribution function

**10.E.1.** Consider rolling a die. The set of sample points is  $\Omega = \{\omega_1, \dots, \omega_6\}$ , where  $\omega_i$  means that we have rolled number  $i$ . Further, consider the  $\sigma$ -field

$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \Omega\}.$$

Find whether the mapping  $X : \Omega \rightarrow \mathbb{R}$  defined by

- i)  $X(\omega_i) = i$  for each  $i \in \{1, 2, 3, 4, 5, 6\}$ ,
- ii)  $X(\omega_1) = X(\omega_2) = -2, X(\omega_3) = X(\omega_4) = X(\omega_5) = X(\omega_6) = 3$

is a random variable with respect to  $\mathcal{A}$ .

**Solution.** First of all, we should make sure that the set  $\mathcal{A}$  really satisfies all axioms of 10.2.2, i. e., that it is a well-defined  $\sigma$ -field. Then, by definition 10.2.10, a random variable is any function  $X : \Omega \rightarrow \mathbb{R}$  such that the preimage of every Borel-measurable set  $B \subset \mathbb{R}$  lies in  $\mathcal{A}$ . As for the first case, consider the interval  $[2, 3]$ . Since  $X^{-1}([2, 3]) = \{\omega_2, \omega_3\} \notin \mathcal{A}$ , we can see that the function  $X$  is not a random variable.

In the second case, we can easily see that  $X$  is a random variable: Consider any interval in  $\mathbb{R}$ . Then, exactly one of the four following occurs: 1) If the interval contains neither  $-2$  nor  $3$ , then the preimage in  $X$  is the empty set. 2) If it contains  $-2$  but not  $3$ , then the preimage is  $\{\omega_1, \omega_2\}$ . 3) On the other hand, if it contains  $3$  but not  $-2$ , then the preimage is  $\{\omega_3, \omega_4, \omega_5, \omega_6\}$ . 4) Finally, if it contains both these numbers, then the preimage is the whole sample space  $\Omega$ . In each case, the preimage lies in the  $\sigma$ -field  $\mathcal{A}$ .  $\square$

### DISTRIBUTION FUNCTION

**Definition.** The *distribution function* or *cumulative distribution function* of a random variable  $X$  is the function  $F_X : \mathbb{R} \rightarrow [0, 1]$  defined for all  $x \in \mathbb{R}$  by

$$F_X(x) = P(X < x).$$

The *distribution function* of a random vector  $(X_1, \dots, X_k)$  is the function  $F_X : \mathbb{R}^k \rightarrow \mathbb{R}$  defined for all vectors  $x = (x_1, \dots, x_k) \in \mathbb{R}^k$  by

$$F_X(x) = P(X_1 < x_1, \dots, X_k < x_k).$$

If it is clear from the context which distribution function is discussed, omit the random variable name and write simply  $F(x)$ .



The following theorem guarantees that, for every random variable, the probability that the value of  $X$  falls into any (fixed) interval (and thus into any Borel set  $B$ ) can be calculated purely from the knowledge of its distribution function.<sup>2</sup>

**10.2.12. Theorem.** For every random variable  $X$ , its distribution function  $F : \mathbb{R} \rightarrow [0, 1]$  has the following properties:

- (1)  $F$  is a non-decreasing function;
- (2)  $F$  has both side-limits at every point  $x \in \mathbb{R}$ , yet these limits may differ;
- (3)  $F$  is left-continuous;
- (4) at the infinite points, the limits of  $F$  are

$$\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1;$$

- (5) the probability of  $X$  taking on the value  $x$  is given by

$$P(X = x) = \lim_{y \rightarrow x+} F(y) - F(x).$$

- (6) The distribution function of a random variable always has only countably many points of discontinuity.

**PROOF.** The proof consists of quite simple and straightforward calculations. In particular, note that the events  $a \leq X < b$  and  $X < a$  are exclusive, so

$$P(a \leq X < b) = P(X < b) - P(X < a) = F(b) - F(a).$$

Hence the first property follows immediately from the definition of probability.

The next two statements follow from the probability of monotone sequences of events, discussed in 10.2.5. Fix a non-increasing sequence of numbers  $r_n > 0$  which converges to 0, and consider the events  $A_n$  given by  $X < x - r_n$ . The union of these events is exactly the event  $A$  given by  $X < x$ . Of course, the event  $A$  does not depend on the choice of the sequence  $r_n$ . By the first proposition of 10.2.5,

$$P(A) = \lim_{n \rightarrow \infty} P(A_n).$$

<sup>2</sup>In literature, the definition with the non-strict inequality  $F(x) = P(X \leq x)$  is often met. In this case, the probability  $P(X = x)$  is also included in  $F_X(x)$ . Then, the distribution function has similar properties as those in 10.2.12, only it is right-continuous instead of left-continuous, etc.

**10.E.2.** Consider a  $\sigma$ -field  $(\Omega, \mathcal{A})$ , where  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$  and

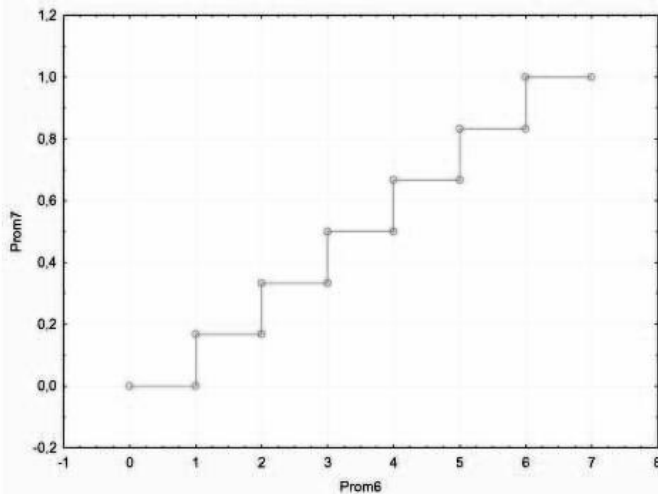
$$\mathcal{A} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4, \omega_5\}, \{\omega_3, \omega_4, \omega_5\}, \Omega\}.$$

Find a mapping  $X : \Omega \rightarrow \mathbb{R}$ , as general as possible, which is a random variable with respect to  $\mathcal{A}$ .

**Solution.** Since the events  $\omega_1, \omega_2$  do not occur individually in  $\mathcal{A}$ , the random variable  $X$  must map them to the same number, i. e.  $X(\omega_1) = X(\omega_2) = a$  for an  $a \in \mathbb{R}$ . For the same reason, we must have  $X(\omega_4) = X(\omega_5) = b$  for a  $b \in \mathbb{R}$ . If an interval contains both  $a$  and  $b$ , then its preimage is  $\{\omega_1, \omega_2, \omega_4, \omega_5\} \in \mathcal{A}$ , which is okay. Clearly, the event  $\omega_3$  may be mapped to an arbitrary  $c \in \mathbb{R}$ . Then, we can easily verify that the  $X$ -preimage of every interval is contained in  $\mathcal{A}$ , i. e.,  $X$  is a random variable with respect to  $\mathcal{A}$ .  $\square$

**10.E.3.** Consider a random variable  $X$  which takes on value  $i$  with probability  $P(X = i) = \frac{1}{6}$ , for each  $i = 1, \dots, 6$ . Find the distribution function  $F_X(x)$  and draw its graph.

**Solution.** By definition 10.2.11, the distribution function is  $F_X(x) = P(X < x)$ . This means that  $F_X(x) = 0$  for  $x < 1$ ,  $F_X(x) = \frac{\lfloor x \rfloor}{6}$  for  $1 \leq x < 6$  (where  $\lfloor x \rfloor$  stands for the floor of  $x$ ), and  $F_X(x) = 1$  for  $x \geq 6$ . The graph looks as follows:



**10.E.4.** An archer keeps shooting at a target until he hits. He has 4 arrows at his disposal. In each attempt, the probability that he hits the target is 0.6. Let  $X$  be the random variable which gives the number of unused arrows. Find the probability mass function and the distribution function of  $X$  and draw their graphs.

**Solution.** Clearly, the probability of  $k$  consecutive misses followed by a hit is equal to  $0.4^k \cdot 0.6$ . Therefore,  $f_X(x) = P(X = x) = 0.4^{3-x} \cdot 0.6$  for  $x \in \{1, 2, 3\}$ . If the archer misses three times, then there will be no arrow left at the

The distribution function is non-decreasing, and thus the left-sided limit equals the supremum. Thus, the left-sided limit  $F_X$  at  $x$  exists and equals  $P(A)$ . This proves one half of proposition (2) as well as all of proposition (3).

Similarly, the above sequence  $r_n$  can be used to define the events  $A_n$  by  $X_n < x + r_n$ . This time, it is a non-increasing sequence  $A_1 \supset A_2 \supset \dots$ , and its intersection is the event  $X \leq x$ . By the second property of 10.2.5,

$$P(A) = \lim_{n \rightarrow \infty} P(A_n) = P(X \leq x),$$

which verifies that the right-sided limit of  $F$  at  $x$  exists. At the same time, property (5) is proved.

The limit values of property (4) can be derived similarly by applying theorem 10.2.5, as shown for the one-sided limits above. In the first case, use the events  $A_n$  given by  $X < r_n$ , for an arbitrary increasing sequence  $r_n \rightarrow \infty$ . Their union is the universal event  $\Omega$ . In the second case, use the events  $A_n$  given by  $X < r_n$ , for any decreasing sequence  $r_n \rightarrow -\infty$ , and their intersection is the null event.

It remains to prove the last statement. As already shown, the discontinuity points of the distribution function are exactly those values  $x$  which the random variable has with non-zero probability, i.e.,  $P(X = x) \neq 0$ . Now, let  $M_n$  denote the set of points  $x$  for which  $P(X = x) > \frac{1}{n}$ . Clearly, the set  $M$  of all discontinuity points equals the union of the sets  $M_n: M = \cup_{n=2}^{\infty} M_n$ . Since the sum of probabilities of mutually exclusive events cannot exceed 1,  $M_n$  can contain no more than  $n - 1$  elements. Thus,  $M$  is a countable union of finite sets, thus it is countable.  $\square$

**10.2.13. Probability measure.** The probability that a random variable has a value lying in an arbitrarily chosen interval can be computed purely from the knowledge of its distribution function. The distribution function  $F_X$  thus defines the entire probability distribution of the random variable  $X$ .

How a particular random variable  $X$  is defined can be ignored.  $X$  can be viewed directly as a probability definition on the  $\sigma$ -algebra of all the Borel sets in  $\mathbb{R}$ .

In this sense, every function  $F : \mathbb{R} \rightarrow \mathbb{R}$  satisfying the first four properties of the latter theorem is a distribution function of a unique random variable. Check the properties of the probability function defined on all intervals this way!

The probability obtained in this way is also called a *probability measure* on  $\mathbb{R}$ . Similarly one deals with probability measures on the algebra of Borel sets in  $\mathbb{R}^k$  in terms of the distribution functions of the random vectors.

In this sense, a random variable or random vector can be considered without any explicit link to a probability space  $(\Omega, \mathcal{A}, P)$ .

**10.2.14. Discrete random variables.** Random variables behave substantially differently according to whether the non-zero probability is “concentrated in isolated points” or it is “continuously distributed” along (a part of) the real axis.



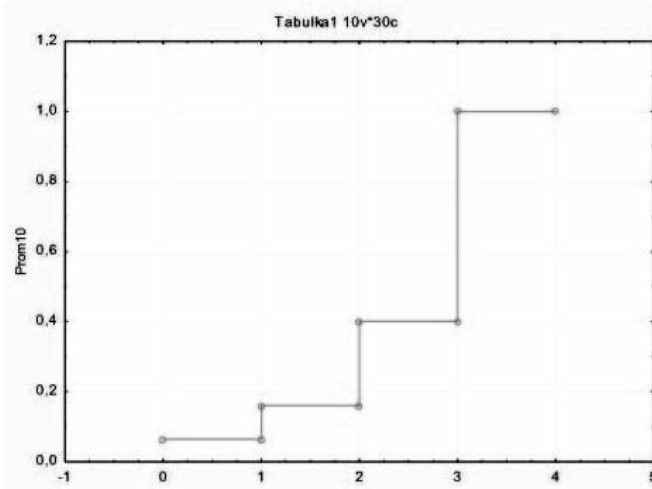
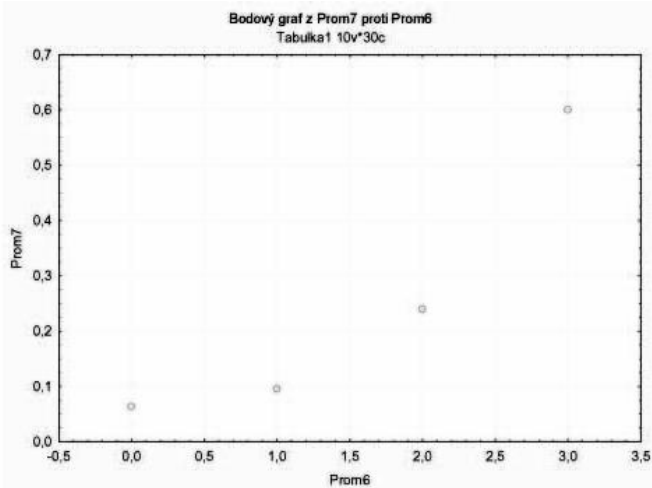


end, no matter whether he hits the last time or not. Thus,  $f_X(0) = P(X = 0) = 0.4^3$ .

By the definition of the distribution function (see 10.2.11), we have

$$F_X(x) = P(X < x) = \begin{cases} 0 & \text{for } x \leq 0, \\ 0.4^3 = 0.064 & \text{for } x \in (0, 1], \\ 0.4^3 + 0.4^2 \cdot 0.6 = 0.16 & \text{for } x \in (1, 2], \\ 0.4^3 + 0.4^2 \cdot 0.6 + 0.4 \cdot 0.6 = 0.4 & \text{for } x \in (2, 3], \\ 1 & \text{for } x > 3. \end{cases}$$

The graphs of the probability mass function and the distribution function are as follows:



□

**10.E.5.** The distribution function of a random variable  $X$  is

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 3 \\ \frac{1}{3}x - 1 & \text{for } 3 < x \leq 6 \\ 1 & \text{for } 6 < x. \end{cases}$$

i) Justify that  $F_X$  is indeed a distribution function.

DISCRETE RANDOM VARIABLES

If a random variable  $X$  assumes only finitely many values  $x_1, x_2, \dots, x_n \in \mathbb{R}$  or countably infinitely many values  $x_1, x_2, \dots$ , it is called a *discrete random variable*.

One can define its *probability mass function*  $f(x)$  by

$$f(x) = \begin{cases} P(X = x_i) & x = x_i \\ 0 & \text{otherwise.} \end{cases}$$

Since the probability is countably additive and the singleton events  $X = x_i$  are mutually exclusive, the sum of all values  $f(x_i)$  is given by either a finite sum or an absolutely convergent series

$$\sum_i f(x_i) = 1.$$

The probability distribution of a random variable  $X$  satisfies

$$P(X \in B) = \sum_{x_i \in B} f(x_i).$$

In particular, the distribution function is of the form

$$F_X(t) = \sum_{x_i < t} f(x_i).$$

Note the the distribution function  $F(x)$  of a discrete random variable is piecewise constant.  $F(x) = 1$  for those  $x$  which are greater than all the  $x_i$ 's.

Every random variable defined on a classical finite probability space is discrete.

**10.2.15. Continuous random variables.** Even if the values of a random variable  $X$  are not discrete, one can proceed similarly. Intuitively, increasing the value of  $x$  infinitesimally by  $dx$ , the density function  $f(x)$  of the random variable  $X$  can be perceived as

$$P(x \leq X < x + dx) = f(x)dx.$$

This means that whenever  $-\infty \leq a \leq b \leq \infty$ , it is required that

$$P(a \leq X < b) = \int_a^b f(x)dx.$$

CONTINUOUS RANDOM VARIABLES

A random variable  $X$  for which there exists a function  $f$  satisfying

$$F_X(b) = \int_{-\infty}^b f(x)dx,$$

is said to be a *continuous random variable*, and the function  $f$  is called its *density function*.

It is convenient to view the random variables as probability measures, cf. 10.2.13. Generally, this means not referring to any other sample space  $\Omega$  on which  $X$  would be defined as a function.  $X$  is represented just by its density or distribution function.

- ii) Find the density of the random variable  $X$ .
- iii) Compute  $P(2 < X < 4)$ .

**Solution.** a) Clearly,  $F_X$  is continuous and non-decreasing. Moreover, we have  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ , as needed.

b) By 10.2.14, the density of a continuous random variable is the derivative of its distribution function. We can see that on the interval  $(3, 6)$ , the density is equal to  $f(x) = \frac{1}{3}$ , while on the intervals  $(-\infty, 3)$  and  $(6, \infty)$ , it is equal to zero. Therefore, the variable  $X$  has uniform distribution, see 10.2.20.

c) We have from the definition of the distribution function that  $P(2 < X < 4) = F_X(4) - F_X(2) = \frac{4}{3} - 1 = \frac{1}{3}$ .  $\square$

**10.E.6.** Consider a random variable  $X$  and a function  $f : \mathbb{R} \rightarrow \mathbb{R}$  given by  $f(x) = \frac{a}{1+x^2}$  for  $x \in \mathbb{R}$ , where  $a$  is a parameter. Suppose that  $f$  is the density of  $X$ . Find

- i) the value of  $a$ ,
- ii) the distribution function of  $X$ ,
- iii)  $P(-1 < X < 1)$ .

**Solution.** a) If the function  $f$  is to be a probability density, then its integral over  $\mathbb{R}$  must be equal to one. This yields the condition

$$1 = \int_{-\infty}^{\infty} \frac{a}{1+x^2} dx = a[\arctg x]_{-\infty}^{\infty} = a\pi.$$

Hence  $a = \frac{1}{\pi}$ .

b) By 10.2.14, the distribution function is given by the following integral:

$$F_X(x) = \int_{-\infty}^x f(t) dt = \frac{1}{\pi} \int_{-\infty}^x \frac{dt}{1+t^2} = \frac{1}{\pi} \arctg x + \frac{1}{2}.$$

c) By b) and the definition of the distribution function, we have

$$P(-1 < X < 1) = F_X(1) - F_X(-1) = \frac{1}{\pi} \cdot \frac{\pi}{4} - \frac{1}{\pi} \cdot \left(-\frac{\pi}{4}\right) = \frac{1}{2}.$$

$\square$

**10.E.7.** The joint probability mass function of a discrete random vector is given by the following table:

Y \ X	2	5	6
1	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{20}$
2	$\frac{1}{10}$	$\frac{1}{20}$	0
3	$\frac{3}{10}$	$\frac{1}{20}$	$\frac{3}{20}$

Find

- i) the marginal distribution and probability mass functions;
- ii) the joint distribution function and draw it in a suitable way;
- iii)  $P(Y > 3X)$ .

**Solution.** a) By 10.2.22, the marginal distribution of the random variable  $X$  is obtained by summing up the joint probability mass function over all possible values of  $Y$  in each row. Similarly, the marginal distribution of  $Y$  is obtained by

Note that the distribution function  $F(x)$  of a continuous random variable  $X$  is always differentiable. Its derivative is the density function of  $X$ , i.e.,  $F'(x) = f(x)$ .

**10.2.16. The general case.** Of course, there are also random variables with mixed behaviour, where some part of the probability is distributed continuously, while there are values that are taken on with non-zero probability. This means, the probability measure of some singletons  $x \in \mathbb{R}$  is non-zero and still  $X$  is not a discrete random variable.



For instance, consider a chaotic lecturer who remains to stand at his laptop with probability  $p$  throughout the entire lecture, but once he decides to move, he happens to be at any position in front of the lecture room with equal probability.

Then, the random variable which corresponds to his position (assume that the desk with the laptop is at position 0 and the lecture room is bounded by the values  $\pm 1$ ) has the following distribution function:

$$F(t) = \begin{cases} 0 & \text{if } t \leq -1 \\ \frac{1-p}{2}(t+1) & \text{if } t \in (-1, 0) \\ p + \frac{1-p}{2}(t+1) & \text{if } t \in [0, 1) \\ 1 & \text{if } t \geq 1. \end{cases}$$

The distribution function of all such variables can be expressed directly using the Riemann-Stieltjes integral

$$F(t) = \int_{-\infty}^t f(x) d(g(x)),$$

developed in subsection 6.3.14 (page 570). In the example above, choose  $f(x) = 1$  and

$$g(x) = \begin{cases} -1 & \text{for } x \leq -1 \\ \frac{1-p}{2}x & \text{for } -1 < x < 0 \\ \frac{1-p}{2}x + p & \text{for } 0 \leq x < 1 \\ \frac{1+p}{2} & \text{for } x \geq 1. \end{cases}$$

This corresponds again to the idea, that the distribution function is equivalent to a probability measure. Thus the measure of any interval is given by integrating its indicator function with respect to this measure. This is what the Riemann-Stieltjes integral achieves.

The Riemann integral corresponds to the choice  $g(x) = x$ . One could add only the jump  $p$  at  $x = 0$  (i.e.  $g(x) = x$  for  $x < 0$ , while  $g(x) = x + p$  otherwise) and leave the constant density  $\frac{1-p}{2}$  to  $f(x)$ , which would be nonzero only on  $[-1, 1]$ . This corresponds to splitting the probability measure into its discrete part (hidden in  $g$ ) and continuous part (expressed by the probability density).

Notice that any distribution function can have only countably many points of discontinuity.

**10.2.17. Basic discrete distributions.** The requirements on the properties of probability distributions of random variables are based on the modeled situations. Here is a list of the simplest discrete distributions.



summing up the entries in each column. Thus, we get the following:

X	1	2	3
$f_X$	$\frac{7}{20}$	$\frac{3}{20}$	$\frac{1}{2}$

and

Y	2	5	6
$f_Y$	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

b) The joint distribution function is at point  $(a, b)$  equal to the sum of all values of the joint probability mass function  $f_{(X,Y)}$  such that  $X \leq a$  and  $Y \leq b$ . This corresponds to values of the subtable whose lower-right corner is  $(a, b)$ . Precisely, the joint distribution function  $F_{(X,Y)}$  looks as follows **is missing ...** and on intervals  $(-\infty, 1) \times \mathbb{R}$  and  $\mathbb{R} \times (-\infty, 2)$ ,  $F_{(X,Y)}$  is clearly zero.

c) Apparently,  $P(Y > 3X) = P(X = 1, Y = 5) + P(X = 1, Y = 6) = \frac{1}{10} + \frac{1}{20} = \frac{3}{20}$  □

**10.E.8.** Find the probability  $P(2X > Y)$  provided the density of the random vector  $(X, Y)$  is given by:

$$f_{(X,Y)}(x, y) = \begin{cases} \frac{1}{6}(4x - y) & \text{for } 1 \leq x \leq 2, 2 \leq y \leq 4, \\ 0 & \text{otherwise.} \end{cases}$$

**Solution.** By definition, we have

$$\begin{aligned} P(2X > Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{2x} f_{(X,Y)}(x, y) dy dx = \\ &= \int_1^2 \int_2^{2x} \frac{1}{6}(4x - y) dy dx = \\ &= \int_1^2 \left[ \frac{2}{3}xy - \frac{1}{12}y^2 \right]_2^{2x} dx = \\ &= \int_2^4 \left( x^2 - \frac{4}{3}x + \frac{1}{3} \right) dx = \\ &= \left[ \frac{1}{3}x^3 - \frac{2}{3}x^2 + \frac{1}{3}x \right]_1^2 = \frac{2}{3}. \end{aligned}$$

□

**10.E.9.** Find the marginal distribution function and the joint and marginal density of the random vector  $(X, Y)$  provided

$$F_{(X,Y)}(x, y) = \begin{cases} 0 & \text{for } x < 0, y < 0 \\ \frac{1}{4}x^2y^2 & \text{for } 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 1 & \text{for } x > 1, y > 2 \end{cases}$$

**Solution.** The density of the random vector  $(X, Y)$  is obtained by differentiation with respect to  $x$  and  $y$ . Thus, for  $0 \leq x \leq 1, 0 \leq y \leq 2$ , we have  $f_{(X,Y)}(x, y) = xy$ , and elsewhere the density is zero. The marginal density of the random variable  $X$  is then

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dy = \int_0^2 xy dy = \left[ \frac{1}{2}xy^2 \right]_0^2 = 2x.$$

Similarly, for  $Y$ , we get  $f_Y(y) = \frac{1}{2}y$ . The marginal distribution functions are

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_0^x 2t dt = x^2$$

DEGENERATE DISTRIBUTION

The distribution which corresponds to a constant random variable  $X = \mu$  is called the *degenerate distribution*  $Dg(\mu)$ .

Its distribution function  $F_X$  and probability mass function  $f_X$  are given by

$$F_X(t) = \begin{cases} 0 & t \leq \mu \\ 1 & t > \mu \end{cases} \quad f_X(t) = \begin{cases} 1 & t = \mu \\ 0 & \text{otherwise.} \end{cases}$$

Here follows a description of an experiment with two possible outcomes called success and failure. If the probability of success is  $p$ , then the probability of failure must be  $1 - p$ .

It is convenient to take the values 0 and 1 for the two possible results.

BERNOULLI DISTRIBUTION

The distribution of a random variable  $X$  which is 0 (failure) with probability  $q = 1 - p$  and 1 (success) with probability  $p$  is called the *Bernoulli distribution*  $A(p)$ .

Its distribution function  $F_X$  and probability mass function  $f_X$  are given by

$$F_X(t) = \begin{cases} 0 & t \leq 0 \\ q & 0 < t \leq 1 \\ 1 & t > 1 \end{cases} \quad f_X(t) = \begin{cases} p & t = 1 \\ q & t = 0 \\ 0 & t \notin \{0, 1\}. \end{cases}$$

Further, consider a random variable  $X$  which corresponds to  $n$  independent experiments described by the Bernoulli distribution, where  $X$  measures the number of successes. Clearly the probability mass function is non-zero exactly at the integers  $t = 0, \dots, n$ , which correspond to the total number of successes in the experiments (the order does not matter).

The probability that  $t$  successes are encountered in  $t$  chosen experiments out of  $n$  is  $p^t(1 - p)^{n-t}$ . It is necessary to sum all the  $\binom{n}{t}$  possibilities. This leads to the binomial distribution of  $X$ :

BINOMIAL DISTRIBUTION

The *binomial distribution*  $Bi(n, p)$  has probability mass function

$$f_X(t) = \begin{cases} \binom{n}{t} p^t (1 - p)^{n-t} & t \in \{0, 1, \dots, n\} \\ 0 & \text{otherwise.} \end{cases}$$

The illustration shows the probability mass functions for  $Bi(50, 0.2)$ , and  $Bi(50, 0.9)$ . The distribution of the probability corresponds to the intuition that most outcomes occur near the value  $np$ :

and

$$F_Y(y) = \int_{-\infty}^y f_Y(t) dt = \int_0^y \frac{1}{2} t dt = \frac{1}{4} y^2. \quad \square$$

**10.E.10.** In a bag, there are 14 balls—4 red, 5 white, and 5 blue ones. We randomly take 6 balls out of the bag (without replacement). Find the distribution of the random vector  $(X, Y)$  where  $X$  stands for the number of red balls taken and  $Y$  for the number of white balls. In addition, find the marginal distributions of  $X$  and  $Y$ . Then, compute  $P(X \leq 3)$ ,  $P(1 \leq Y \leq 4)$ .

**Solution.** The value of the probability mass function at point  $(x, y)$  is defined as the probability  $P(X = x, Y = y)$ , i. e. the probability of taking  $x$  red balls and  $y$  white balls. The number of ways how to take  $x$  red balls is  $\binom{4}{x}$ ; for  $y$  white balls, it is  $\binom{5}{y}$ ; and the remaining  $6 - x - y$  blue balls can be selected in  $\binom{5}{6-x-y}$  ways. Altogether, there are  $\binom{4}{x} \binom{5}{y} \binom{5}{6-x-y}$  possibilities. The values of this expression for all  $x, y$  are in the following table.

$x \backslash y$	0	1	2	3	4	5	$\sum_X$
0	0	5	50	100	50	5	210
1	4	100	400	400	100	4	1008
2	30	300	600	300	30	0	1260
3	40	200	200	40	0	0	480
4	10	25	10	0	0	0	45
$\sum_Y$	84	630	1260	840	180	9	3003

The values in the last column and row are the sums over all values of  $y$  and  $x$ , respectively. Then, the values of the probability mass function are obtained after dividing by the number of all possibilities how to take the 6 balls, i. e.  $\binom{14}{6} = 3003$ . The marginal distributions of  $X$  and  $Y$  correspond to the last column and row, respectively.

The probability  $P(X \leq 3)$  can be calculated easily from the marginal distribution of  $X$ :

$$P(X \leq 3) = F_X(3) = \frac{1}{3003} (210 + 1008 + 1260 + 480) = 0.985.$$

Similarly, for the probability  $P(1 \leq Y \leq 4)$ , we have

$$\begin{aligned} P(1 \leq Y \leq 4) &= F_Y(4) - F_Y(1) = \\ &= \frac{1}{3003} (630 + 1260 + 840 + 180) = 0.969. \end{aligned} \quad \square$$

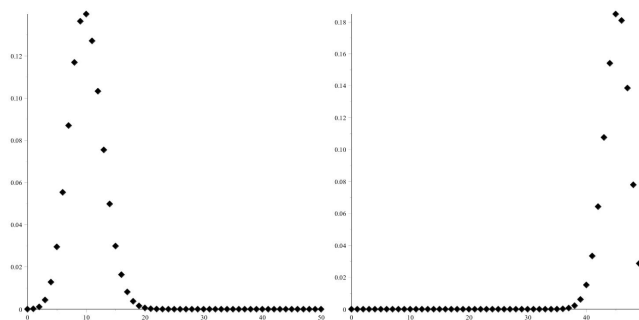
**10.E.11.** The density of a random vector  $(X, Y, Z)$  is

$$f(x, y, z) = \begin{cases} c(x + y + z) & \text{for } 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of the parameter  $c$  as well as the distribution function of the vector, and compute  $P(0 \leq X \leq \frac{1}{2}, 0 \leq Y \leq \frac{1}{2}, 0 \leq Z \leq \frac{1}{2})$ .

**Solution.** The integral of the density over the entire space must be equal to one. This gives us

$$\begin{aligned} 1 &= \int_0^1 \int_0^1 \int_0^1 c(x + y + z) dz dy dx = c \int_0^1 \int_0^1 (x + y + \frac{1}{2}) dy dx \\ &= c \int_0^1 (x + 1) dx = \frac{3}{2} c. \end{aligned}$$



Next, consider distributions similar to the Bernoulli process referred to in 10.2.1. Consider independent experiments with the Bernoulli distribution  $A(p)$ , as in the case of the binomial distribution, and fix a positive integer  $r$ . Repeat the experiment until  $r$  successes occur.

The random variable  $X$  is defined as the number of failures before the  $r$ -th success. In the case of  $r = 1$ , it is exactly the example from 10.2.1. The event  $X = k$  occurs if and only if there are exactly  $r - 1$  successes in the first  $k + r - 1$  experiments and the  $(k + r)$ -th experiment also ends with a success. Thus the following probability mass function is arrived at:

#### GEOMETRIC DISTRIBUTION

The random variable  $X$  which corresponds to the number of failures before reaching the  $r$ -th success has probability distribution

$$P(X = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \quad k = 0, 1, 2, \dots$$

This is called the *negative binomial distribution*. In the case of  $r = 1$ , it is the *geometric distribution*.

Often the same definition is used with the successes and failures interchanged. This results in the same formula for the probability mass function with  $p$  and  $1 - p$  interchanged.

The geometric distribution appears in physics in connection with Einstein–Bose statistics.

**10.2.18. Poisson distribution.** In practice, the binomial distribution often leads to further model problems.

Consider the situation that  $r$  (mutually indistinguishable) objects are to be divided into  $n$  (distinguishable) boxes, and each object is equally probable (i.e., has probability  $1/n$ ) to fall into any of the boxes.

The random variable which describes the number  $X$  of objects in one fixed box can be described as follows: The admissible values are  $X = k$ , where  $k = 0, \dots, r$ , and the individual probabilities are

$$P(X = k) = \binom{r}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k} = \binom{r}{k} \frac{(n-1)^{r-k}}{n^r}.$$

Thus, the distribution of  $X$  is of the type  $Bi(r, 1/n)$ .

Such a variable can be encountered, for example, when describing a physical system with a huge number of gas molecules. The boxes represent small volumes of the space.

Hence,  $c = \frac{2}{3}$ . By definition, the distribution function is equal to

$$\begin{aligned} F_X(x, y, z) &= \frac{2}{3} \int_0^x \int_0^y \int_0^z (r + s + t) dt ds dr = \\ &= \frac{2}{3} \int_0^x \int_0^y (rz + sz + \frac{1}{2}z^2) ds dr = \frac{2}{3} \int_0^x (ryz + \frac{1}{2}y^2z + \frac{1}{2}z^2y) dy = \\ &= \frac{2}{3} (\frac{1}{2}x^2zy + \frac{1}{2}y^2zx + \frac{1}{2}z^2yx) = \frac{2}{3} (x^2zy + y^2zx + z^2yx), \end{aligned}$$

so the wanted probability is

$$P(0 \leq X \leq \frac{1}{2}, 0 \leq Y \leq \frac{1}{2}, 0 \leq Z \leq \frac{1}{2}) = F(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}) = \frac{1}{16}. \quad \square$$

**10.E.12.** Find the value of the parameter  $a$  so that the function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 1 \\ a \ln(x) & \text{for } 1 < x < 2 \\ 0 & \text{for } 2 \leq x \end{cases}$$

would be the probability density of a random variable.

**Solution.** We know that the condition for the function to be a density is

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Thus, we have to calculate  $\int \ln(x) dx$ :

$$\int \ln(x) dx = x \ln(x) - \int 1 dx = x \ln(x) - x = x(\ln(x) - 1).$$

Altogether,

$$\int_{-\infty}^{\infty} f(x) dx = \int_1^2 a \ln(x) dx = a[x(\ln(x) - 1)]_1^2 = a(2 \ln(2) - 1),$$

so  $a = \frac{1}{2 \ln(2) - 1}$ . □

**10.E.13.** A child has become lost in a forest whose shape is that of a regular hexagon. Suppose that the probability that the child happens to be in a given part of the forest is directly proportional to the size of that part, but independent of its position in the forest.

- What is the probability distribution of the distance of the child from a given side (extended to a straight line) of the forest?
- What is the probability distribution of the distance of the child from the closest side of the forest?

**Solution.**

- Let  $a$  be the length of the sides of the hexagon (forest). Then, the probability distribution satisfies

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{4}{9a^2}x + \frac{2}{3\sqrt{3}a} & \text{for } 0 < x \leq \frac{1}{2}\sqrt{3}a \\ -\frac{4}{9a^2}x + \frac{2}{\sqrt{3}a} & \text{for } \frac{1}{2}\sqrt{3}a \leq x \leq \sqrt{3}a \\ 0 & \text{for } x > \sqrt{3}a \end{cases},$$

as for the first question.

- First, let us compute the distribution function  $F$  of the wanted random variable  $X$  that corresponds to the distance of the child from the closest side. The distance can

Observe the distribution of the molecules. Then, the behaviour of  $X_n$  as the number  $n$  of boxes as well as the number  $r_n$  of objects increases so that their ratio  $r_n/n = \lambda$  remains constant is of interest. In other words, every box is to contain (approximately) the same number  $\lambda$  of elements, on average.

We are interested in the asymptotic behaviour of the variables  $X_n$  as  $n \rightarrow \infty$ . Letting  $\lim_{n \rightarrow \infty} r_n/n = \lambda$ , the standard procedure (with details to be added – take it as a challenge to recall the methods from the analysis of univariate functions!) leads to:

$$\begin{aligned} \lim_{n \rightarrow \infty} P(X_n = k) &= \lim_{n \rightarrow \infty} \binom{r_n}{k} \frac{(n-1)^{r_n-k}}{n^{r_n}} \\ &= \lim_{n \rightarrow \infty} \frac{r_n(r_n-1) \dots (r_n-k+1)}{(n-1)^k} \frac{1}{k!} \left(1 - \frac{1}{n}\right)^{r_n} \\ &= \frac{\lambda^k}{k!} \lim_{n \rightarrow \infty} \left(1 + \frac{-r_n}{r_n}\right)^{r_n} = \frac{\lambda^k}{k!} \lim_{m \rightarrow \infty} \left(1 + \frac{-\lambda}{m}\right)^m \\ &= \frac{\lambda^k}{k!} e^{-\lambda}, \end{aligned}$$

since the functions  $(1+x/n)^n$  converge uniformly to the function  $e^x$  on every bounded interval in  $\mathbb{R}$ .

POISSON DISTRIBUTION

The *Poisson distribution*  $Po(\lambda)$  describes the random variables with probability mass function

$$f_X(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & k \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases}$$

Of course,

$$\sum_{k=0}^{\infty} f_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda + \lambda} = 1.$$

As seen above, this discrete distribution  $Po(\lambda)$  with an arbitrary  $\lambda > 0$  (distributed into infinitely many points) is a good approximation of the binomial distributions  $Bi(n, \lambda/n)$ , for large values of  $n$ .

**10.2.19. Two examples.** Besides the physical model mentioned above, such a behaviour can be encountered when observing occurrences of events in a space with constant expected density in a unit volume. Observing bacteria under a microscope, when the bacteria are expected to occur in any part of the image with the same probability, provides an example. If the “mean density of occurrence” in a unit area is  $\lambda$  and the whole region is divided into  $n$  identical parts, then the occurrence of  $k$  events in a fixed part is modeled by a random variable  $X$  with the Poisson distribution. When diagnosing in practice, such an observation allows us to compute the total number of bacteria with a relatively good accuracy from the actual numbers in only several randomly chosen samples.



be anywhere in the interval  $I = \langle 0, \frac{\sqrt{3}}{2}a \rangle$ . Then, for  $y \in I$ , we have

$$F(y) = P[X < y] = \frac{\frac{\sqrt{3}}{4}a^2 - \frac{(\frac{\sqrt{3}}{2}a - y)^2}{\frac{3}{4}a^2} \frac{\sqrt{3}}{4}a^2}{\frac{\sqrt{3}}{4}a^2} = 1 - \frac{4(\frac{\sqrt{3}}{2}a - y)^2}{3a^2}$$

Altogether,

$$F(y) = \begin{cases} 0 & \text{for } y \leq 0 \\ 1 - \frac{4(\frac{\sqrt{3}}{2}a - y)^2}{3a^2} & \text{for } y \in \langle 0, \frac{\sqrt{3}}{2}a \rangle \\ 1 & \text{for } y \geq \frac{\sqrt{3}}{2}a \end{cases}$$

Thus, the density, being the derivative of the distribution function, satisfies:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{8(\frac{\sqrt{3}}{2}a - y)}{3a^2} & \text{for } y \in \langle 0, \frac{\sqrt{3}}{2}a \rangle \\ 0 & \text{for } y \geq \frac{\sqrt{3}}{2}a \end{cases}$$

□

**10.E.14.** Let a random variable  $X$  have uniform distribution on an interval  $\langle 0, r \rangle$ . Find the distribution function and probability density of the volume of the ball whose radius is equal to  $X$ .

**Solution.** First, we find the distribution function  $F$  (for  $0 < d < \frac{4}{3}\pi r^3$ )

$$F(d) = P\left[\frac{4}{3}\pi X^3 \leq d\right] = P\left[X \leq \sqrt[3]{\frac{3d}{4\pi}}\right] = \frac{\sqrt[3]{\frac{3d}{4\pi}}}{r}$$

Altogether,

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sqrt[3]{\frac{3}{4\pi r^3}} x^{\frac{1}{3}} & \text{for } 0 < x < \frac{4}{3}\pi r^3 \\ 1 & \text{for } x \geq \frac{4}{3}\pi r^3 \end{cases}$$

Differentiating this, we obtain the density:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sqrt[3]{\frac{1}{36\pi r^3}} x^{-\frac{2}{3}} & \text{for } 0 < x < \frac{4}{3}\pi r^3 \\ 0 & \text{for } x \geq \frac{4}{3}\pi r^3 \end{cases}$$

□

**10.E.15.** Find the value(s) of the parameter  $a \in \mathbb{R}$  so that the function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ ax^2 & \text{for } 0 < x < 3 \\ 0 & \text{for } x \geq 3 \end{cases}$$

defines the probability density of a random variable  $X$ . Then, find its distribution function, probability density, and the expected value of the volume of the cube whose edge-length has probability density determined by  $f$ .

**Solution.** Simply,  $a = \frac{1}{9}$ . Thus, the distribution function of the random variable  $X$  is  $F_X(t) = \frac{1}{27}t^3$  for  $t \in (0, 3)$ , zero for smaller values of  $t$ , and one for greater. Let  $Z = X^3$  denote the random variable corresponding to the volume of the considered cube. It lies in the interval  $(0, 27)$ . Thus, for

The second example is more challenging. We describe events which occur randomly at time  $t \geq 0$ . Here, the probability of an occurrence in the following small time period of length  $h$  does not depend on what had happened before and equals the same value  $h\lambda$  for a fixed  $\lambda > 0$ . At the same time, the probability that the event occurs more than once in a given time period is small.

Let  $X_t$  denote the random variable which corresponds to the number of occurrences of the examined event in the interval  $[0, t)$ . The requirements are expressed infinitesimally. We want:

- the probability of exactly one event in each time period of length  $h$  equals  $h\lambda + \alpha(h)$ , where the function  $\alpha(h)$  satisfies  $\lim_{h \rightarrow 0_+} \frac{\alpha(h)}{h} = 0$ ;
- the probability  $\beta(h)$  of more than one event occurring in a time period of length  $h$  to satisfy  $\lim_{h \rightarrow 0_+} \frac{\beta(h)}{h} = 0$ ;
- the events  $X_t = j$  and  $X_{t+h} - X_t = k$  to be independent for all  $j, k \in \mathbb{N}$  and  $t, h > 0$ .

Use the notation  $p_k(t) = P(X_t = k)$ ,  $k \in \mathbb{N}$ , and set the initial conditions  $p_0(0) = 1$  and  $p_k(0) = 0$  for  $k > 0$ . Compute directly

$$p_0(t+h) = p_0(t)P(X_{t+h} - X_t = 0) = p_0(t)(1 - h\lambda - \alpha(h) - \beta(h))$$

and similarly,

$$\begin{aligned} p_k(t+h) &= P(X_t = k, X_{t+h} - X_t = 0) \\ &\quad + P(X_t = k-1, X_{t+h} - X_t = 1) \\ &\quad + P(X_t \leq k-2, X_{t+h} = k) \\ &= p_k(t)P(X_{t+h} - X_t = 0) + p_{k-1}P(X_{t+h} - X_t = 1) \\ &\quad + \sum_{i=0}^{k-2} P(X_t = i, X_{t+h} - X_t = k-i) \\ &= p_k(t)(1 - h\lambda - \alpha(h) - \beta(h)) + p_{k-1}(t)(h\lambda + \alpha(h)) \\ &\quad + \sum_{i=0}^{k-2} p_i(t)P(X_{t+h} - X_t = k-i). \end{aligned}$$

Hence (similar to in 6.1.12, page 520, the symbol  $o(h)$  is written for expressions which, when divided by  $h$ , approach zero as  $h \rightarrow 0_+$ )

$$\begin{aligned} \frac{p_0(t+h) - p_0(t)}{h} &= -\lambda p_0(t) + \frac{1}{h} o(h) \\ \frac{p_k(t+h) - p_k(t)}{h} &= -\lambda p_k(t) + \lambda p_{k-1}(t) + \frac{1}{h} o(h). \end{aligned}$$

Letting  $h \rightarrow 0_+$ , an (infinite!) system of ordinary differential equations is obtained:

$$\begin{aligned} p'_0(t) &= -\lambda p_0(t), & p_0(0) &= 1 \\ p'_k(t) &= -\lambda p_k(t) + \lambda p_{k-1}(t), & p_k(0) &= 0 \end{aligned}$$

for all  $t > 0$  and  $k \in \mathbb{N}$ , with an initial condition.

The first equation has a unique solution

$$p_0(t) = e^{-\lambda t},$$

$t \in (0, 27)$  and the distribution function  $F_Z$  of the random variable  $Z$ , we can write  $F_Z(t) = P[Z < t] = P[X^3 < t] = P[X < \sqrt[3]{t}] = F_X(\sqrt[3]{t}) = \frac{1}{27}t$ . Then, the density is  $f_Z(t) = \frac{1}{27}$  on the interval  $(0, 27)$  and zero elsewhere. Since this is the uniform distribution on the given interval, the expected value is equal to 13.5.  $\square$

**10.E.16.** Find the value(s) of the parameter  $a \in \mathbb{R}$  so that the function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ ax & \text{for } 0 < x < 3 \\ 0 & \text{for } x \geq 3 \end{cases}$$

defines the probability density of a random variable  $X$ . Then, find its distribution function, probability density, and the expected value of the area of the square whose side-length has probability density determined by  $f$ .

**Solution.** We proceed similarly as in the previous example. Again, we can easily find that  $a = \frac{2}{9}$ . Thus, the distribution function of the random variable  $X$  is  $F_X(t) = \frac{1}{9}t^2$  for  $t \in (0, 3)$ , zero for smaller values of  $t$ , and one for greater. Let  $Z = X^3$  denote the random variable corresponding to the area of the considered square. It lies in the interval  $(0, 9)$ . Thus, for  $t \in (0, 9)$  and the distribution function  $F_Z$  of the random variable  $Z$ , we can write  $F_Z(t) = P[Z < t] = P[X^2 < t] = P[X < \sqrt{t}] = F_X(\sqrt{t}) = \frac{1}{9}t$ . Then, the density is  $f_Z(t) = \frac{1}{9}$  on the interval  $(0, 9)$  and zero elsewhere. Since this is the uniform distribution on the given interval, the expected value is equal to 4.5.  $\square$

**10.E.17.** Find the value(s) of the parameter  $a \in \mathbb{R}$  so that the function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ ax^2 & \text{for } 0 < x < 2 \\ 0 & \text{for } x \geq 2 \end{cases}$$

defines the probability density of a random variable  $X$ . Then, find its distribution function, probability density, and the expected value of the volume of the cube whose edge-length has probability density determined by  $f$ .  $\circ$

**10.E.18.** We randomly cut a line segment of length  $l$  into two pieces. Find the distribution function and the density of the area of the rectangle whose side-lengths are equal to the obtained pieces.

**Solution.** Let us compute the distribution function: Let  $X$  denote the random variable with uniform distribution on the interval  $\langle 0, l \rangle$ , which corresponds to the length of one of the pieces (then, the length of the other piece is  $l - X$ ). The area  $S = x(l - x)$  of the rectangle, for  $x \in \langle 0, l \rangle$ , can lie anywhere in the interval  $\langle 0, l^2/4 \rangle$ . Setting  $d \in \langle 0, l^2/4 \rangle$ , we can write

$$F(d) = P[S \leq d] = P[X(l - X) \leq d]$$

Thus, we are looking for those values of  $x$  for which  $x(l - x) \leq d$ , which is a quadratic inequality. The roots of the corresponding quadratic equation are  $\frac{l - \sqrt{l^2 - 4d}}{2}$  and  $\frac{l + \sqrt{l^2 - 4d}}{2}$ .

which can be immediately substituted into the second equation. This leads to

$$p_1(t) = \lambda t e^{-\lambda t}.$$

A trivial induction argument shows that the system has a unique solution

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad t > 0, \quad k \in \mathbb{N}.$$

It is thus verified that for every process which satisfies the three properties above, the random variable  $X_t$  which corresponds to the number of occurrences in the time period  $[0, t)$  has distribution  $\text{Po}(\lambda t)$ .

In practice, these processes are connected with the failure rate of machines.

**10.2.20. Continuous distributions.** The simplest example of a continuous distribution is the uniform distribution of the probability throughout a fixed interval. This is also a good illustration of the fact that a simply formulated requirement does not leave many free choices in the definition. Now, the probability of  $X$  taking on a value inside an interval which is included in the sample interval  $(a, b) \subset \mathbb{R}$  is required to be dependent only on the length of the interval, but not on its actual position. This means that the density  $f_X$  of the random variable  $X$  should be constant and the value of this constant is given by the requirement  $P(a \leq X < b) = 1$ .



UNIFORM DISTRIBUTION

For any real numbers  $a, b, -\infty < a < b < \infty$ , define the density and distribution function as follows:

$$f_X(t) = \begin{cases} 0 & t \leq a \\ \frac{1}{b-a} & t \in (a, b) \\ 0 & t \geq b, \end{cases} \quad F_X(t) = \begin{cases} 0 & t \leq a \\ \frac{t-a}{b-a} & t \in (a, b) \\ 1 & t \geq b. \end{cases}$$

Here, the random variable  $X$  has *uniform distribution*.

The next distribution is similar to the discrete Poisson distribution. Suppose the occurrence of a random event is observed such that its occurrences in non-overlapping intervals are independent. Thus, if  $p(t)$  is the probability of the event not occurring during an interval of length  $t$ , then of necessity  $p(t + s) = p(t)p(s)$  for all  $t, s > 0$ . Moreover, assume that  $p$  is differentiable and  $p(0) = 1$ .

Then,  $\ln p(t + s) = \ln p(t) + \ln p(s)$ . Letting  $s \rightarrow 0_+$  (and applying l'Hospital's rule),

$$\begin{aligned} (\ln(p))'(t) &= \lim_{s \rightarrow 0_+} \frac{\ln p(t + s) - \ln p(t)}{s} \\ &= \lim_{s \rightarrow 0_+} \frac{(\ln p(s))'}{1} = \frac{p'(0)}{p(0)} = p'(0). \end{aligned}$$

Thus,  $p'(0) = -\lambda \in \mathbb{R}$  (Note:  $\lambda > 0$ , and  $p'(0)$  cannot be positive as  $p(0) = 1$ ).

Then,  $p(t)$  satisfies  $\ln p(t) = -\lambda t + C$ . The initial condition leads to the only solution

$$p(t) = e^{-\lambda t}.$$

The inequality is satisfied by exactly those values of  $x$  which lie outside this interval. Therefore,

$$P[X(l - X) \leq d] = P[X \in \langle 0, l \rangle \setminus \left( \frac{l - \sqrt{l^2 - 4d}}{2}, \frac{l + \sqrt{l^2 - 4d}}{2} \right)] \\ = \frac{l - \sqrt{l^2 - 4d}}{l} = 1 - \frac{\sqrt{l^2 - 4d}}{l}$$

Altogether,

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - \frac{\sqrt{l^2 - 4x}}{l} & \text{for } 0 \leq x \leq \frac{l^2}{4} \\ 1 & \text{for } x > \frac{l^2}{4} \end{cases}$$

The density is obtained by differentiation:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{2}{l\sqrt{l^2 - 4x}} & \text{for } 0 \leq x \leq \frac{l^2}{4} \\ 0 & \text{for } x > \frac{l^2}{4} \end{cases}$$

□

**10.E.19.** Nezávislé náhodné veličiny  $X$  a  $Y$  mají následující hustoty pravděpodobnosti:

$$f_X(t) = \begin{cases} 0 & \text{for } t \leq 0, \\ 1 & \text{for } 0 < t < 1, \\ 0 & \text{for } 1 \leq t, \end{cases} \quad f_Y(t) = \begin{cases} 0 & \text{for } t \leq 0, \\ 2t & \text{for } 0 < t < 1, \\ 0 & \text{for } 1 \leq t. \end{cases}$$

Určete distribuční funkci náhodné veličiny udávající obsah obdélníka o stranách  $X$  a  $Y$ .

**Solution.**

$$F_Y(t) = \begin{cases} 0 & \text{for } t \leq 0 \\ 2t - t^2 & \text{for } 0 < t < 1 \\ 1 & \text{for } 1 \leq t \end{cases}$$

□

**10.E.20.** Let  $X, Y$  be independent random variables, where  $X$  has uniform distribution on the interval  $(0, 2)$  and  $Y$  is given by its density function:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 2x & \text{for } 0 < x < 1 \\ 0 & \text{for } x \geq 1. \end{cases}$$

Find the probability that  $Y$  is less than  $X^2$ .

**Solution.** Since  $X$  and  $Y$  are independent random variables, the joint density  $f_{(X,Y)} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  of the variable  $(X, Y)$  is given by the densities  $f_X$  and  $f_Y$  of the individual random variables. Thus, we have

$$f_{(X,Y)}(u, v) = \begin{cases} f_X(u) \cdot f_Y(v) = \frac{1}{2} \cdot 2v = v & \text{for } (u, v) \in (0, 2) \times (0, 1), \\ 0 & \text{otherwise.} \end{cases}$$

Now, consider the random variable  $X$  which corresponds to a (random) moment when the event occurs for the first time. Apparently, the distribution function of  $X$  is given by

$$F_X(t) = 1 - p(t) = \begin{cases} 1 - e^{-\lambda t} & t > 0 \\ 0 & t \leq 0. \end{cases}$$

This function has the desired properties: It has values between zero and one, it is increasing and it has the required behaviour at  $\pm\infty$ . The density of this random variable can be obtained by differentiation of the distribution function.

#### EXPONENTIAL DISTRIBUTION

The distribution corresponding to the continuous random variable  $X$  with density

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & t > 0 \\ 0 & t \leq 0. \end{cases}$$

is called the *exponential distribution*  $\text{ex}(\lambda)$ .

The exponential distribution belongs to the more general family of important distributions with the densities of the form

$$cx^{a-1} e^{-bx}$$

for  $x > 0$ , with given constants  $a > 0, b > 0$ , while the constant  $c$  is to be computed. The following expression is required to equal one:

$$\int_0^\infty cx^{a-1} e^{-bx} dx = \int_0^\infty c \left(\frac{t}{b}\right)^{a-1} e^{-t} \frac{1}{b} dt = \frac{c}{b^a} \Gamma(a).$$

$\Gamma$  is the famous transcendental function providing the analytic extension of the factorial function, discussed in 6.2.17 on the page 546.

#### GAMMA DISTRIBUTION

The distribution whose density is zero for  $x \leq 0$ , while for  $x > 0$ . It is given by

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx},$$

called the *gamma distribution*  $\Gamma(a, b)$  with parameters  $a > 0, b > 0$ .

Thus, the exponential distribution is the special case of this one for the value  $a = 1$ .

**10.2.21. Normal distribution.** Recall the binomial distribution. If the success rate  $p$  is left constant, but the number  $n$  of experiments is increased, the probability mass function keeps its shape (although the scale changes). As  $n$  increases, the values of the probability mass function merges into a curve that should correspond to the density of a continuous distribution which is a good approximation for  $\text{Bi}(n, p)$  for large values of  $n$ .



Recall the smooth function  $y = e^{-x^2/2}$ , mentioned in subsection 6.1.9 (page 516) as an appropriate tool for the construction of functions which are smooth but not analytic. The



Then, the wanted probability  $P$  is the integral of the density  $f_{(X,Y)}$  over the part  $O$  of the plane where  $Y < X^2$ :

$$P = \iint_O f_{(X,Y)} dx dy = 1 - \iint_{\mathbb{R}^2 \setminus O} f_{(X,Y)} dx dy = 1 - \int_0^1 \int_{x^2}^1 y dy dx = \frac{3}{5}.$$

□

**10.E.21.** Let  $X, Y$  be independent random variables, where  $X$  has density function

$$f_1(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 2x & \text{for } 0 < x < 1 \\ 0 & \text{for } x \geq 1, \end{cases}$$

and  $Y$  has density function

$$f_2(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{x}{2} & \text{for } 0 < x < 2 \\ 0 & \text{for } x \geq 2. \end{cases}$$

Find the probability that  $Y$  is greater than  $X^2$ . ○

**10.E.22.** Let  $X, Y$  be independent random variables, where  $X$  has density function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{2x}{9} & \text{for } 0 < x < 3 \\ 0 & \text{for } x \geq 3, \end{cases}$$

and  $Y$  has density function

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{x}{2} & \text{for } 0 < x < 2 \\ 0 & \text{for } x \geq 2. \end{cases}$$

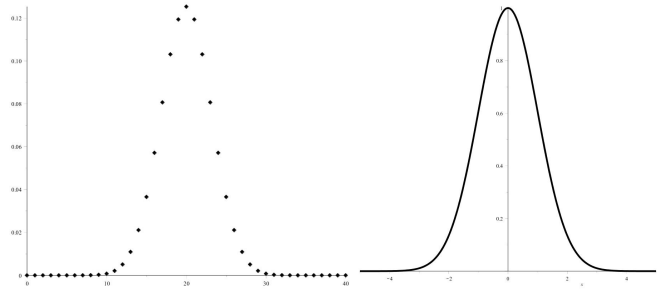
Find the probability that  $Y$  is greater than  $X^3$ . ○

### F. Expected value, correlation

Compute the expected value and variance of the binomial distribution.

**Solution.** The direct calculation from the definitions is a nice exercise on combinatorics. We prove this statement using the properties of the expected value and variance. Using the definition of the binomial distribution (see 10.2.17), we can view the random variable  $X \sim \text{Bi}(n, p)$  as the sum  $X = \sum_{k=1}^n Y_k$ , where  $Y_1, \dots, Y_n \sim A(p)$  are independent random variables saying whether the  $k$ -th experiment was successful. Clearly, the Bernoulli distribution has expected value  $E Y_i = p$ , hence by theorem 10.2.29, we have  $E X = \sum_{k=1}^n E Y_k = np$ . Similarly, we compute  $E(Y_k^2) = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$ , so  $\text{var } Y_k = E(Y_k^2) - (E Y_k)^2 = p - p^2$ . By theorem 10.2.33, we have  $\text{var } X = \sum_{k=1}^n \text{var } Y_k = np(1 - p)$ . □

illustration compares this curve (in the right hand part) to the values of  $\text{Bi}(40, 0.5)$ .



This suggests looking for a convenient continuous distribution whose density would be given by a suitably adjusted variation of this function.

The function  $e^{-x^2/2}$  is everywhere positive, so it suffices to compute  $\int_{-\infty}^{\infty} e^{-x^2/2} dx$ . If this results in a finite value, just multiply the function by its reciprocal value. Unfortunately, this integral cannot be computed in terms of elementary functions. Luckily, multidimensional integration and Fubini's theorem can be used. Transform to polar coordinates, to obtain

$$\begin{aligned} & \left( \int_{-\infty}^{\infty} e^{-x^2/2} dx \right) \left( \int_{-\infty}^{\infty} e^{-y^2/2} dy \right) \\ &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2} dx dy \\ &= \int_0^{\infty} \int_0^{2\pi} e^{-(r^2)/2} r dr d\theta \\ &= 2\pi \end{aligned}$$

(cf. the notes at the end of subsection 8.2.5, verify that the integrated function satisfies the conditions given there, and compute that thoroughly!). Hence the integral results in  $\sqrt{2\pi}$ , so the function  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$  is a well-defined density of a random variable.

### NORMAL DISTRIBUTION

The distribution of the random variable  $Z$  with density

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

is called the (standard) *normal distribution*  $N(0, 1)$ . The corresponding distribution function

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$

cannot be expressed in terms of elementary functions.

It is called the *Gaussian function* and the graph of  $\varphi(x)$  is often called the *Gaussian curve*.

So far, the correct density which approximates the binomial distribution is not found. The diagram that compares the probability function of the binomial distribution to the Gaussian curve shows that the position of the maximum must be moved as well as an application of shrinkage or stretch to the curve horizontally. The first goal is easily reached by constant

**10.F.1.** An archer shoots five arrows at a target. Each time, the probability he hits is 0.6, and the individual results are independent. Let  $X$  be the random variable which corresponds to the number of hits. Determine its distribution and find its expected value and variance.

**Solution.** Clearly, the shots are independent experiments with the Bernoulli distribution  $A(\frac{3}{5})$ . Thus, by the definition of the binomial distribution, we have  $X \sim \text{Bi}(5, \frac{3}{5})$ . By **F**, the expected value and variance of  $\text{Bi}(n, p)$  are equal to  $np$  and  $np(1-p)$ , respectively, which gives  $E X = 3$  and  $\text{var } X = \frac{6}{5}$  for our case.  $\square$

**10.F.2.** Consider the discrete random variable  $X$  which takes on the values  $k = 0, 1, 2, 3, \dots$ , each with probability  $P(X = k) = p(1-p)^k$  (geometric distribution). Find  $E X$  (the expected number of failures before the first success) and  $\text{var } X$ .

**Solution.** Using the definition of the expected value and the formula for summing the derivative of a geometric series, we calculate

$$\begin{aligned} E X &= \sum_{k=0}^{\infty} k p (1-p)^k = p(1-p) \sum_{k=0}^{\infty} k (1-p)^{k-1} = \\ &= p(1-p) \frac{1}{p^2} = \frac{1-p}{p}. \end{aligned}$$

Similarly, using the formula for summing the second derivative of a geometric series, we compute

$$E(X^2) = \sum_{k=0}^{\infty} k^2 p (1-p)^k = \frac{(1-p)(2-p)}{p^2},$$

hence the variance is  $\text{var } X = E(X^2) - (E X)^2 = \frac{1-p}{p^2}$ .  $\square$

**10.F.3.** A random variable  $X$  is defined by its density  $f_X(x) = \frac{3}{x^4}$  for  $x \in (1, \infty)$  and  $f_X(x) = 0$  elsewhere. Find its distribution function, expected value, and variance.

**Solution.** By the definition of the distribution function, we have, for  $x \in (1, \infty)$ ,

$$F_X(x) = \int_1^x \frac{3}{t^4} dt = \left[ -\frac{1}{t^3} \right]_1^x = 1 - \frac{1}{x^3}.$$

The expected value of  $X$  is equal to

$$E X = \int_1^{\infty} \frac{3}{x^3} dx = \left[ -\frac{3}{2x^2} \right]_1^{\infty} = \frac{3}{2}$$

and the expected value of  $X^2$  is

$$E(X^2) = \int_1^{\infty} \frac{3}{x^2} dx = \left[ -\frac{3}{x} \right]_1^{\infty} = 3.$$

Therefore,  $\text{var } X = 3 - (\frac{3}{2})^2 = \frac{3}{4}$ .

shift  $\mu$  of the variable  $z$ , while scaling the difference  $x - \mu$  by coefficient  $\sigma > 0$  does the rest. Thus, there are two real parameters  $\mu$  and  $\sigma > 0$  and the density function is of the form:

$$g_{\mu, \sigma}(x) = e^{-(x-\mu)^2 / (2\sigma^2)}.$$

Simple variable substitution leads to

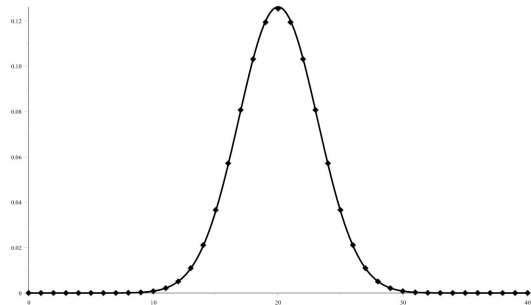
$$\int_{-\infty}^{\infty} e^{-(x-\mu)^2 / (2\sigma^2)} dx = \sqrt{2\pi} \sigma.$$

Thus there is an entire two-parametric class of densities

$$\varphi_{\mu, \sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

of random variables. The corresponding distributions are denoted by  $N(\mu, \sigma^2)$ .

We return to the asymptotic closeness of the normal and binomial distributions for  $n \rightarrow \infty$  after creating suitable tools. The following illustration reveals, how well this works. The discrete values correspond to  $\text{Bi}(40, 0.5)$ , while the curve depicts the density of  $N(20, 10)$ .



**10.2.22. Distributions of random vectors.** As for the scalar random variables, one defines the distribution functions and the density or the probability mass function for continuous and discrete random vectors. There are joint probability mass functions and densities.

For two discrete random variables, i.e. a discrete vector  $(X, Y)$  of random variables, define their *(joint) probability mass function*

$$f(x, y) = \begin{cases} P(X = x_i \wedge Y = y_j) & x = x_i, y = y_j \\ 0 & \text{otherwise.} \end{cases}$$

A random vector  $(X, Y)$  is called continuous, if its distribution function is defined as for continuous random variables. This means, for all  $a, b \in \mathbb{R}$ ,

$$\begin{aligned} F(a, b) &= P(X < a, Y < b) = \\ &= \int_{-\infty}^b \int_{-\infty}^a f(x, y) dx dy, \end{aligned}$$

and the function  $f(x, y)$  is called the *(joint) density* of the random vector  $(X, Y)$ .  $\square$

**10.F.4.** A random variable  $X$  is defined by its density  $f_X(x) = \cos x$  for  $x \in \langle 0, \frac{\pi}{2} \rangle$  and  $f_X(x) = 0$  elsewhere. Find its expected value, variance, and median.

**Solution.** Using the definition and integration by parts, we get

$$E X = \int_0^{\frac{\pi}{2}} x \cos x dx = [x \sin x + \cos x]_0^{\frac{\pi}{2}} = \frac{\pi}{2} - 1.$$

Using double integration by parts, we obtain

$$\begin{aligned} E(X^2) &= \int_0^{\frac{\pi}{2}} x^2 \cos x dx = \\ &= [x^2 \sin x + 2x \cos x - 2 \sin x]_0^{\frac{\pi}{2}} = \left(\frac{\pi}{2}\right)^2 - 2, \end{aligned}$$

so the variance is equal to  $\text{var } X = (\frac{\pi}{2})^2 - 2 - (\frac{\pi}{2} - 1)^2 = \pi - 3$ . By definition, the distribution function is equal to  $F_X(x) = \int_0^x \cos t dt = \sin x$ , and the median is  $F^{-1}(0.5) = \frac{\pi}{6}$ .  $\square$

**10.F.5.** A random variable  $X$  is defined by its density  $f_X(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$ , and  $f_X(x) = 0$  elsewhere (the so-called exponential distribution;  $\lambda > 0$  is a fixed parameter). Find its expected value, variance, mode (the real number where the density reaches its maximum), and median.

**Solution.** Using the definition and integration by parts, we get

$$E X = \int_0^{\infty} x \lambda e^{-\lambda x} dx = \left[ -x e^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \right]_0^{\infty} = \frac{1}{\lambda},$$

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \\ &= \left[ -x^2 e^{-\lambda x} - 2x \frac{1}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x} \right]_0^{\infty} = \frac{2}{\lambda^2}, \end{aligned}$$

hence  $\text{var } X = E(X^2) - (E X)^2 = \frac{1}{\lambda^2}$ . Since  $F'_X(x) = -\lambda^2 e^{-\lambda x} < 0$ , the density keeps decreasing. Therefore, its maximum is at zero. By definition, we have

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x},$$

so the median is equal to  $F^{-1}(0.5) = -\frac{1}{\lambda} \ln(\frac{1}{2}) = \frac{\ln 2}{\lambda}$ .  $\square$

**10.F.6.** The joint probability mass function of a discrete random vector  $(X_1, X_2)$  is defined by  $\pi(0, -1) = c$ ,  $\pi(1, 0) = \pi(1, 1) = \pi(2, 1) = 2c$ ,  $\pi(2, 0) = 3c$  and zero elsewhere. Find the parameter  $c$  and compute the covariance  $\text{cov}(X_1, X_2)$ .

**Solution.** If  $\pi$  is to be a probability mass function, then the sum of its values over the entire domain must be equal to 1, i. e.,

$$\sum_{i,j} \pi(i, j) = c + 3.2c + 3c = 10c = 1,$$

so  $c = \frac{1}{10}$ . The probability mass function  $\pi_1$  of  $X_1$  is given by the sum of the joint function over all possible values of  $X_2$ , i. e.,  $\pi_1(i) = \sum_j \pi(i, j)$ . Thus, we have  $\pi_1(0) = c$ ,  $\pi_1(1) = 4c$ ,  $\pi_1(2) = 5c$  and zero elsewhere. Similarly, for

For a general continuous random vector  $X = (X_1, \dots, X_n)$ , define

$$\begin{aligned} F(a_1, \dots, a_n) &= P(X_1 < a_1, \dots, X_n < a_n) = \\ &= \int_{-\infty}^{a_n} \dots \int_{-\infty}^{a_1} f(x_1, \dots, x_n) dx_1 \dots dx_n, \end{aligned}$$

and similarly for discrete random vectors with more components.

A random vector  $(X, Y)$  with both  $X$  and  $Y$  continuous is not always a continuous vector in the above sense. For example, taking a continuous variable  $X$ , the random vector  $(X, 2X)$  is neither continuous nor discrete, since the entire probability mass is concentrated along the line  $y = 2x$  in the plane, but not in individual points.

The *marginal distribution* for one of the variables can be obtained by summation or integration over the others.

For instance, in the case of a discrete random vector  $(X, Y)$ , the events  $(X = x_i, Y = y_j)$  for all possible values  $x_i$  and  $y_j$  with non-zero probabilities for  $X$  and  $Y$ , respectively, form an exhaustive collection of events for the vector  $(X, Y)$ . Thus

$$P(X = x_i) = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j),$$

which relates the *marginal probability distribution* of the random variable  $X$  to the joint probability distribution of the random vector  $(X, Y)$ . In the case of continuous random vectors, proceed similarly using integrals instead of sums.

**10.2.23. Stochastic independence.** It is known from



subsection 10.2.3 what (in)dependence means for events. Random variables  $X_1, \dots, X_n$  are (*stochastically*) *independent* if and only if for any  $a_i \in \mathbb{R}$ , the events  $X_1 < a_1, \dots, X_n < a_n$  are independent. In view of the definition of the distribution function  $F$  of the random vector  $(X_1, \dots, X_n)$ , this is equivalent to

$$F(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n),$$

where  $F_{X_i}$  are the distribution functions of the individual components.

It follows that the events corresponding to  $X_k \in I_k$  for arbitrarily chosen intervals  $I_k$  is also independent. The probability of  $X_1 \in [a, b)$  and simultaneously  $X_i \in (-\infty, c_i)$  for the other components is  $F(b, c_2, \dots, c_n) - F(a, c_2, \dots, c_n) = (F_{X_1}(b) - F_{X_1}(a))F_{X_2}(c_2) \dots F_{X_n}(c_n)$ , and so on. The densities and probability mass functions behave well too:

**Proposition.** For any random vector  $(X_1, \dots, X_n)$ , the following two conditions are equivalent:

- The random variables  $X_1, \dots, X_n$  are stochastically independent.
- The joint distribution function  $F$  of the random vector  $(X_1, \dots, X_n)$  is the product of the marginal distribution functions  $F_{X_i}$  of the individual components.

the probability mass function  $\pi_2$  of  $X_2$ , we get  $\pi_2(-1) = c$ ,  $\pi_2(0) = 5c$ ,  $\pi_2(1) = 4c$  and zero elsewhere. Hence,  $E X_1 = \sum_i i\pi_1(i) = 14c = 1.4$  and  $E X_2 = \sum_j j\pi_2(j) = 3c = 0.3$ . By the definition of the covariance, we have

$$\text{cov}(X_1, X_2) = \sum_{i,j} (i-1, 4)(j-0, 3)\pi(i, j) = 0.18.$$

□

**10.F.7.** In many scientific fields, the behavior of a random variable which is bounded onto an interval is modeled using the so-called beta-distribution. That is a continuous distribution is defined by its density on the interval  $[0, 1]$ :

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where  $\alpha, \beta$  are fixed parameters, chosen suitably for description of the given random variable, and  $B(\alpha, \beta)$  is a normalizing constant, guaranteeing that the integral of  $f_X(x)$  over  $[0, 1]$  is equal to zero. Find its a) mode, b) expected value, and c) variance.

**Solution.** a) By definition, the mode is the value where  $f_X(x)$  reaches its maximum. Thus, let us look at its stationary points. We can easily calculate that the equation  $f'_X(x) = 0$  is equivalent to

$$(\alpha-1)(1-x) - x(\beta-1) = 0,$$

and this is satisfied for  $x = \frac{\alpha-1}{\alpha+\beta-2}$ . Since  $f_X(0) = f_X(1) = 0$  and the function is positive, it must be the wanted maximum.

b) By definition, we have

$$E X = \frac{1}{B(\alpha, \beta)} \int_0^1 x^\alpha(1-x)^{\beta-1} dx.$$

Integrating by parts, we get

$$E X = -\frac{1}{B(\alpha, \beta)\beta} [x^\alpha(1-x)^\beta]_0^1 + \frac{\alpha}{B(\alpha, \beta)\beta} \int_0^1 x^{\alpha-1}(1-x)^\beta dx.$$

Clearly, the first term is equal to zero. Refining the second one, we obtain

$$E X = \frac{\alpha}{B(\alpha, \beta)\beta} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx - \frac{\alpha}{B(\alpha, \beta)\beta} \int_0^1 x^\alpha(1-x)^{\beta-1} dx.$$

Now, the integral in the first term is, thanks to the normalization, equal to  $B(\alpha, \beta)$ , and the second integral is the expected value, too. Thus, the above equation can be written as

$$E X = \frac{\alpha}{\beta} - \frac{\alpha}{\beta} E X.$$

Hence it immediately follows that  $E X = \frac{\alpha}{\alpha+\beta}$ .

c) In order to compute the variance, we must calculate

$$E X^2 = \frac{1}{B(\alpha, \beta)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1} dx.$$

This integral can be computed similarly as in b):

$$E X^2 = \frac{\alpha+1}{B(\alpha, \beta)\beta} \int_0^1 x^\alpha(1-x)^\beta dx = \frac{\alpha+1}{\beta} E X - \frac{\alpha+1}{\beta} E X^2$$

Moreover, if all  $X_i$  are discrete random variables, then they are independent if and only if the joint probability mass function  $f$  of the random vector  $(X_1, \dots, X_n)$  is the product of the marginal probability mass functions  $f_{X_i}$  of the individual components.

Similarly, if all  $X_i$  are continuous random variables, then they are independent if and only if the joint density function  $f$  of the random vector  $(X_1, \dots, X_n)$  exists and it is the product of the marginal density functions  $f_{X_i}$  of the individual components.

In particular, any random vector with independent continuous components is again a continuous random vector.

**PROOF.** Many of the claims are already verified. The only nontrivial implication left is the one assuming the product formula for the joint distribution function and deriving the claim on the probability function or the density. The argument for  $n = 2$  is shown below, the general case is analogous.

Consider first two discrete independent random variables  $X, Y$ . Then  $f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) = f_X(x_i)f_Y(y_j)$ . The joint distribution function is

$$\begin{aligned} F_{X,Y}(x, y) &= \sum_{x_i < x} \sum_{y_j < y} f_X(x_i)f_Y(y_j) \\ &= \left( \sum_{x_i < x} f_X(x_i) \right) \left( \sum_{y_j < y} f_Y(y_j) \right), \end{aligned}$$

which is equivalent to  $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ .

Similarly, assuming that the joint distribution function  $F_{X,Y}$  is the product of the distributions functions of two continuous random variables  $X, Y$ , then its mixed partial derivatives exist. Thus is set:

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) \\ &= \frac{\partial^2}{\partial x \partial y} F_X(x)F_Y(y) \\ &= f_X(x)f_Y(y), \end{aligned}$$

which is the requested joint density function for  $F_{X,Y}$ .

All the other implications are either direct consequences of the definitions or are obvious. □

**10.2.24. Example.** Consider a simple example which illustrates that it is not a good idea to view a random vector only as a pair of random variables. Consider stochastic properties of a random vector  $(X, Y)$  which has continuous uniform distribution on the unit disc in the plane  $\mathbb{R}^2$ , centered at the origin. Then, its (joint) density function is

$$f(x, y) = \begin{cases} \frac{1}{\pi} & x^2 + y^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The components  $X$  and  $Y$  of this random vector (in the usual Euclidean coordinates) cannot be independent random variables: For instance, note that the probability of  $(X, Y)$  falling outside the unit disc but inside the square with vertices at

Hence,  $E X^2 = \frac{(\alpha+1) E X}{\alpha+\beta+1}$ . Substituting the expected value, we obtain

$$\frac{\text{var } X}{(\alpha + \beta + 1)(\alpha + \beta)^2} = \frac{E X^2 - (E X)^2}{\alpha \beta} \quad \square$$

**10.F.8.** We toss three coins. Let  $X$  denote the total number of heads on the first and second coins, and  $Y$  denote the total number of heads on the second and third coins.

**Solution.** First of all, we build the table describing the joint probability mass function of the discrete random vector  $(X, Y)$ , whence we can get the probability distributions of the variables we will need:

	Y	0	1	2
X				
0		$\frac{1}{8}$	$\frac{1}{8}$	0
1		$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
2		0	$\frac{1}{8}$	$\frac{1}{8}$

The discrete variables  $X$  and  $Y$  have the same probability distribution: they take on the value 0 with probability  $1/4$ , 1 with  $1/2$ , and 2 with  $1/4$ . (Of course, we could have come to this even without the table.) The variable  $XY$  takes on the values 0, 1, 2, 4 with probabilities  $3/8, 1/4, 1/4, 1/8$ , respectively. Now, we compute the expected values of the variables  $X, X^2, Y, Y^2, XY$ :

$$\begin{aligned} E(X) &= E(Y) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1 \\ E(X^2) &= E(Y^2) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = \frac{3}{2} \\ E(XY) &= 0 \cdot \frac{3}{8} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} = \frac{5}{4} \end{aligned}$$

Thus, we have

$$\begin{aligned} \sigma^2(X) &= \zeta^2(Y) = E(X^2) - [E(X)]^2 = \frac{1}{2} \\ \text{cov}(X, Y) &= E(XY) - E(X)E(Y) = \frac{1}{4} \end{aligned}$$

Altogether,

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)} = \frac{1}{2} \quad \square$$

**10.F.9.** Consider random variables  $U, V$ , defined by their joint probability mass function ( $U$  takes on 1 or 2,  $V$  takes on 1, 2 or 3):

	V		
U	1	2	3
1	0.1	0.2	0.3
2	0.2	0.1	0.1

$(\pm 1, \pm 1)$  is zero, while the marginal distribution functions are non-zero for the values  $|x| < 1$  and  $|y| < 1$ .

Expressing this random vector in polar coordinates  $(R, \Phi)$ ,

$$P(R < r_0, \Phi < \varphi_0) = \int_0^{r_0} \int_0^{\varphi_0} \frac{1}{\pi} r \, d\varphi \, dr = \frac{1}{2\pi} \varphi_0 r_0^2.$$

The joint density of the vector  $(R, \Phi)$  is thus  $f(r, \varphi) = \frac{r}{\pi}$  for  $0 < r \leq 1, 0 < \varphi \leq 2\pi$ , and it is zero otherwise. The marginal densities are

$$\begin{aligned} f_R(r) &= \int_0^{2\pi} \frac{r}{\pi} \, d\varphi = 2r, \quad \text{if } 0 < r \leq 1, \\ f_\Phi(\varphi) &= \int_0^1 \frac{r}{\pi} \, dr = \frac{1}{2\pi}, \quad \text{if } 0 < \varphi \leq 2\pi, \end{aligned}$$

and zero otherwise. Therefore, the random variables  $R$  and  $\Phi$  are independent. This is a very important feature in mathematical statistics.

**10.2.25. Functions of random variables.** In practice, random



vectors are encountered in two quite different roles. Firstly, we can observe several random variables which describe less or more related events. As an example, examine various numerical parameters connected to individual students (their results in particular courses, weight, height, age, annual income, etc.). In this case, tools are needed which allow the examination of differences or dependencies between these variables.

We can examine only one of the parameters on a large collection of objects and select only a small number  $n$  of them. This procedure is described by an  $n$ -dimensional vector  $(X_1, \dots, X_n)$  where all the random variables  $X_k$  have the same probability distribution. In this case, we are more interested in the quantities that correspond to statistical numerical characteristics discussed in the previous part of this chapter.

Both cases can be dealt with using one simple concept. Instead of the given random variable or random vector, consider functions of those.

This is a useful tool even in the case of one random variable. As an example, consider the random variable  $X$  with uniform distribution over the interval  $[1, 2] \subset \mathbb{R}$  giving the length of the side of a square, asking for the random variable  $Y = X^2$  describing the area of such a square. The problem is to see the stochastic behaviour of  $Y$  in terms of the known parameters of  $X$ .

The obvious technical condition on  $\psi$  is to guarantee that  $Y = \psi \circ X$  is again a random variable according to the definition. This means the preimage  $\psi^{-1}$  of a Borel-measurable set should be again a Borel-measurable set.

Elementary arguments reveal that  $\psi^{-1}(A \setminus B) = \psi^{-1}(A) \setminus \psi^{-1}(B)$  and  $\cup_{i \in I} \psi^{-1}(A_i) = \psi^{-1}(\cup_{i \in I} A_i)$ , for subsets  $A, B, A_i \in \mathbb{R}$ . Since each open subset in  $\mathbb{R}^n$  is a countable union of intervals, and the pre-images  $\psi^{-1}(U)$  are open for continuous functions  $\psi$  and open sets  $U$ , the continuous function  $\psi$  always satisfies the condition.

Find the marginal distributions of both variables, their expected values, variances, and correlation coefficient. ○

**10.F.10.** Find the expected value and variance of the random variable  $X^2$  provided  $X$  has uniform distribution on the interval  $\langle -1, 1 \rangle$ . ○

**10.F.11.** We roll two dice. Let  $X$  denote how many times we got an even number, and  $Y$  denote how many times we got an odd number. Find their correlation coefficient. ○

**10.F.12.** Consider random variables  $U, V$ , defined by their joint probability mass function ( $U$  takes on 1 or 2,  $V$  takes on 1, 2 or 3):

	V		
U	1	2	3
1	0.1	0.1	0.4
2	0.2	0.1	0.1

Find the marginal distributions of both variables, their expected values, variances, and correlation coefficient. ○

**G. Transformations of random variables**

Consider a continuous function of a random variable  $Y = \psi(X)$ . If the transformation  $\psi$  is increasing, then the resulting distribution function satisfies

$$F_Y(y) = P(Y \leq y) = P(\psi(X) \leq y) = P(X \leq \psi^{-1}(y)) = F_X(\psi^{-1}(y)),$$

where  $F_X$  is the distribution function of  $X$  (analogously for decreasing  $\psi$ ). Thus, the density of the transformed random variable  $Y$  satisfies

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(\psi^{-1}(y)) \left| \frac{d\psi^{-1}(y)}{dy} \right|.$$

Applying the rule for transformation of coordinates, we can compute the expected value of  $Y$  as

$$EY = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx,$$

ans similarly for the variance of  $Y$ .

**10.G.1.** Consider a random variable  $X$  with density  $f(x)$ . Find the density of the random variable  $Y$  defined by

- i)  $Y = e^X, x \geq 0,$
- ii)  $Y = \sqrt{X}, x > 0,$
- iii)  $Y = \ln X, x > 0,$
- iv)  $Y = \frac{1}{X}, x > 0.$

**Solution.** We can simply apply the formula for the density of a transformed random variable, which yields a)  $f_Y(y) = f(\ln y) \frac{1}{y}$ , b)  $f_Y(y) = 2f(y^2)y$ , c)  $f_Y(y) = f(e^y)e^y$ , d)  $f_Y(y) = f(1/y) \frac{1}{y^2}$ . □

In the sequel, we restrict ourselves to this case.

FUNCTIONS OF RANDOM VARIABLES AND VECTORS

For a continuous function  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  and a random variable  $X$ , there is also the random variable  $Y = \psi(X)$ .  $Y$  is said to be a *function of the random variable  $X$* .

In the case of a random vector  $(X_1, \dots, X_n)$  and a continuous function  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ , we talk about a function  $Y = \psi(X_1, \dots, X_n)$  of the *random vector*.

It is useful to know whether independent random variables remain independent after transformations. The answer and its verification are simple:

**Proposition.** Consider two independent random variables  $X$  and  $Y$  and two functions  $g$  and  $h$  such that  $U = g(X)$ ,  $V = h(X)$  are random variables. Then  $U$  and  $V$  are independent, too.

**PROOF.** For fixed reals  $u, v$ , write

$$A_u = \{x; g(x) < u\} \quad B_v = \{y; h(y) < v\}.$$

Then the joint distribution function for the vector  $(U, V)$  is

$$\begin{aligned} F_{U,V}(u, v) &= P(U < u, V < v) = P(X \in A_u, Y \in B_v) \\ &= P(X \in A_u)P(Y \in B_v) \\ &= P(U < u)P(V < v) = F_U(u)F_V(v). \end{aligned}$$

Thus, the transformed random variables are stochastically independent, as expected. □

**10.2.26. Affine transformations and sums.** The simplest function (except constants) is an affine dependency

$$\psi(X) = a + bX$$

with constants  $a, b \in \mathbb{R}, b \neq 0$ .

If  $f_X(x)$  is the probability mass function of a random variable with discrete distribution, it is easily computed that

$$(1) \quad f_{\psi(X)}(y) = P(\psi(X) = y) = \sum_{\psi(x_i)=y} f(x_i).$$

Thus, in the case of the affine dependency  $Y = a + bX$ , the probability mass function is non-zero exactly at the points  $y_i = ax_i + b$ .

As an example of a function of a random vector  $X = (X_1, \dots, X_n)$ , consider the sum of  $n$  independent random variables with the Bernoulli distribution  $X_i \sim A(p)$ . Of course, this leads just to the binomial distribution  $\text{Bi}(n, p)$ . The above formula for  $f_{\psi(X)}(y)$  reveals the already known probability function for  $Y = X_1 + \dots + X_n$ . Only  $y \in \{0, \dots, n\}$  can be reached. Collect all the possibilities of summing  $y$  ones, when each of them appears with probability  $p^y(1-p)^{n-y}$ .

Similarly, proceed with continuous random variables. The two parameter family  $Y = \mu + \sigma Z$  is met already, where

**10.G.2.** Consider a random variable  $X$  which has uniform distribution on the interval  $(-\frac{\pi}{2}, \frac{\pi}{2})$ . Find the density of  $X$  as well as the densities of transformed variables  $Y = \sin X, Z = \tan X$ .

**Solution.** Since the length of the interval where  $X$  is non-zero is  $\pi$ , the density of  $X$  is  $f_X(x) = \frac{1}{\pi}$  for  $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$  and zero elsewhere. Applying the formula for the density of a transformed random variable and the derivatives of elementary functions, we get

$$f_Y(y) = f_X(\arcsin(y)) \arcsin'(y) = \frac{1}{\pi\sqrt{1-y^2}}$$

and

$$f_Z(z) = f_X(\arctan(z)) \arctan'(z) = \frac{1}{\pi(1+z^2)}.$$

□

**10.G.3.** Consider a random variable  $X$  whose density is  $\cos x$  for  $x \in (0, \frac{\pi}{2})$  and zero elsewhere. Find the density of the random variable  $Y = X^2$  and calculate  $EY$  and  $\text{var } Y$ .

**Solution.** Applying the formula for the density of a transformed random variable, we get

$$f_Y(y) = f_X(\sqrt{y})(\sqrt{y})' = \frac{1}{2\sqrt{y}} \cos x.$$

It is simpler to compute the expected value and variance of  $Y$  directly from the density of  $X$ . We have  $EY = \int_{-\infty}^{\infty} x^2 f_X(x) dx$ . Thus,

$$EY = \int_0^{\frac{\pi}{2}} x^2 \cos x dx = [x^2 \sin x + 2x \cos x - 2 \sin x]_0^{\frac{\pi}{2}} = \frac{\pi^2 - 8}{4}.$$

The integral was computed by parts. Applying this method again, we obtain

$$E(Y^2) = \int_0^{\frac{\pi}{2}} x^4 \cos x dx = [(x^4 - 12x^2 + 24) \sin x + 4(x^3 - 6x) \cos x]_0^{\frac{\pi}{2}}.$$

Hence,  $E(Y^2) = (\frac{\pi}{2})^4 - 12(\frac{\pi}{2})^2 + 24$ , so  $\text{var } Y = \frac{\pi^4}{16} - 3\pi^2 + 24 - \frac{\pi^4 - 16\pi^2 + 64}{16} = 20 - 2\pi^2$ .

□

**10.G.4.** Let  $X$  be a random variable which takes on values 0 and 1, each with probability  $\frac{1}{2}$ . Similarly, let  $Y$  take on the values  $-1$  and 1, each with probability  $\frac{1}{2}$ . Show that the random variables  $X$  and  $Z = XY$  are uncorrelated, yet dependent. Give an example of two continuous random variables with this property.

**Solution.** First of all, we compute the expected values of our random variables:  $E X = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$ ,  $E Z = E(XY) = 0 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$ . As for the expected value of their product, we have  $E(XZ) = E(X^2 Y) = 1 \cdot \frac{1}{4} + (-1) \cdot \frac{1}{4} = 0$ . By theorem 10.2.33, the covariance is equal to  $\text{cov}(X, Z) = 0 - \frac{1}{2} \cdot 0 = 0$ . Thus, the variables  $X$  and  $Y$  are uncorrelated. At the same time, the conditional probability  $P(Z = 1 | X = 0)$  is clearly zero, i. e., we also have  $P(Z = 1, X = 0) = 0$ , while  $P(Z = 1) = \frac{1}{4}$  and  $P(X = 0) = \frac{1}{2}$ , so  $P(Z = 1) \cdot P(X = 0) = \frac{1}{8} \neq 0$ . We can

$Z \sim N(0, 1)$  in 10.2.21. This is verified easily.

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(\mu + \sigma Z < y) \\ &= \Phi\left(\frac{y - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y - \mu}{\sigma}} e^{-z^2/2} dz \\ &= \int_{-\infty}^y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}} dx, \end{aligned}$$

where the substitution  $x = \mu + \sigma z$  is used in the last step. This is exactly what is wanted.

More generally, the above formula (1) has the straightforward analog for the density of  $Y = \psi(X)$  for a continuous  $X$  in the case when  $\psi$  has got non-zero derivative (thus  $\psi$  is invertible).

$$(2) \quad f_Y(y) = \int_{-\infty}^y |\psi'(\psi^{-1}(y))|^{-1} f_X(\psi^{-1}(y)) dy.$$

Check the formula yourself! (Start with the case when the derivative  $\psi'$  is always positive.)

It is more complicated with more general sums of independent random variables. Consider two such continuous random variables  $X$  and  $Y$  with densities  $f_X$  and  $f_Y$ , respectively. The distribution function of the random variable  $V = X + Y$  is computed directly (exploit the independence of  $X$  and  $Y$  write the joint density of  $(X, Y)$  as product):



$$\begin{aligned} F_V(u) &= \int_{x+y < u} f_X(x) f_Y(y) dx dy = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{u-x} f_X(x) f_Y(y) dy dx \\ &= \int_{-\infty}^u \left( \int_{-\infty}^{\infty} f_X(x) f_Y(v-x) dx \right) dv, \end{aligned}$$

where the substitution  $v = x + y$  is used together with the Fubini theorem. Thus, the joint density of the sum of two independent random variables is just the convolution of their densities

$$(3) \quad f_V = f_X * f_Y,$$

already met in subsection 7.2.2 (page 656). Similarly, there is a discrete convolution of probability mass functions in the case of discrete random variables.

In the seventh chapter, we viewed the convolution as a kind of blurry picture of one of the functions with the help of the kernel expressed by the other. This should be the right intuition for the density of the sum of independent random variables as well. Of course, this also suggests that the convolution must be symmetric in the arguments.

**10.2.27. Numerical characteristics.** When examining some values (of a measurement, for example) from the statistical point of view, important numerical characteristics like the arithmetic mean and the standard deviation are looked for. Now we introduce similar characteristics for random variables and



see that  $P(Z = 1) \cdot P(X = 0) \neq P(Z = 1, X = 0)$ , which means that  $X$  and  $Z$  are dependent.

It can be easily verified from the corresponding definitions that if  $X$  is any random variable with zero expected value, finite second moment and zero third moment, then  $X$  and  $Y = X^2$  are dependent, but uncorrelated.  $\square$

**H. Inequalities and limit theorems**

Markov’s inequality provides a rough estimate of the behavior of a non-negative random variable if we know nothing more than its expected value. In exact words, for any non-negative random variable  $X$  and any  $a > 0$ , it holds that  $P(X \geq a) \leq \frac{EX}{a}$ .

**10.H.1.** Consider a non-negative random variable  $X$  with expected value  $\mu$ . With no further information about  $X$ , bound  $P(X > 3\mu)$ . Then, compute  $P(X > 3\mu)$  if you know that  $X \sim \text{Ex}(\frac{1}{\mu})$ .

**Solution.** If the non-negative random variable  $X$  does not take zero with probability 1, then its expected value  $\mu$  is positive. Therefore, the wanted probability can be bounded using Markov’s inequality as

$$P(X \geq 3\mu) \leq \frac{\mu}{3\mu} = \frac{1}{3}.$$

If we know that  $X \sim \text{Ex}(\frac{1}{\mu})$ , then

$$P(X > 3\mu) = 1 - P(X \leq 3\mu) = 1 - F(3\mu),$$

where  $F$  is the distribution function of the exponential distribution. By definition, this is

$$F(x) = \int_0^x \frac{1}{\mu} e^{-\frac{t}{\mu}} dt = \left[-e^{-\frac{t}{\mu}}\right]_0^x = 1 - e^{-\frac{x}{\mu}}.$$

Hence,  $P(X > 3\mu) = \frac{1}{e^3}$ .  $\square$

**10.H.2.** At a particular place, the average speed of wind is 20 kilometers per hour.

- Regardless of the distribution of the speed as a random variable, bound the probability that in a given observation, the speed does not exceed 60 km/h.
- Find the interval in which the speed lies with probability at least 0.9 if you know that the standard deviation is  $\sigma = 1$  km/h.

**Solution.** Let  $X$  denote the random variable that corresponds to the speed. In the first case, we can only use Markov’s inequality, leading to

$$P(X \leq 60) = 1 - P(X \geq 60) \geq 1 - \frac{20}{60} = \frac{2}{3}.$$

In the second case, we know the variance (or standard deviation) of the speed, so we can use Chebyshev’s inequality (see 10.2.32):

$$0.9 \leq P(|X - 20| < x) = 1 - P(|X - 20| \geq x) \leq 1 - \frac{1}{x^2}.$$

Hence,  $x \geq \sqrt{10} \approx 3.2$ . Thus, the wanted interval is (16.8 km/h, 23.2 km/h).  $\square$

random vectors. The first one is an analogy of the arithmetic mean.

EXPECTED VALUE

For any random variable  $X$ , define its *expected value*  $EX$  by

$$EX = \begin{cases} \sum_i x_i f_X(x_i) & \text{for a discrete variable} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{for a continuous variable,} \end{cases}$$

provided the sum or integral converges absolutely. If not, the random variable  $X$  is said to have no expected value.

The *expected value of a random vector* is simply the vector of expected values of the individual components.

The expected value can also be expressed directly for functions  $Y = \psi(X)$  of a random variable or vector  $X$ . Recall that we consider only those functions  $\psi$  for which  $Y$  is again a random variable.

In the discrete case, compute

$$\begin{aligned} EY &= \sum_j y_j P(Y = y_j) \\ &= \sum_j y_j \sum_{\psi(x_i)=y_j} P(X = x_j) \\ &= \sum_i \psi(x_i) P(X = x_i), \end{aligned}$$

provided the sum converges absolutely. Of course, it is not guaranteed that a function of a random variable which has expected value also has it.

Similarly, the expected value of a function of a continuous random variable is :

$$E\psi(X) = \int_{-\infty}^{\infty} \psi(x) f_X(x) dx,$$

provided the integral converges absolutely.

Note that the random variable  $Y = \psi(X)$  does not have to be continuous even if the original variable  $X$  is. Nevertheless, if  $\psi$  is a differentiable monotone function with non-zero derivative, it is an easy exercise to verify that the definition of  $E\psi(X)$  coincides with  $EY$ . We do not go into further details here.

Shortly, in the part devoted to statistics, it is shown that the expected value has a direct connection with the arithmetic mean of the corresponding vector of values.

**10.2.28. St. Petersburg paradox.** We return to the example used as motivation for the need of discrete random variables in subsection 10.2.1. Reformulate the model as potential rules for a casino. This results in a good example of a situation where the expected value of the examined random variable does not exist at all.

The gambler pays an initial amount  $C$  and then keeps tossing a coin until it comes up heads. Denoting the number of tosses he has made by  $T$ , he wins  $2^T$ . The problem is to determine a “reasonable value” for the initial amount  $C$ . If  $X$





**10.H.3.** Each yogurt of an undisclosed company contains a photo of one of 26 ice-hockey world champions. Suppose the players are distributed uniformly at random. How many yogurts must Vera buy if she wants the probability of getting at least 5 photos of Jaromír Jágr to be at least 0.95?

**Solution.** Let  $X$  denote the random variable that corresponds to the number of obtained photos of Jaromír Jágr (parametrized by the number  $n$  of yogurts bought). Clearly,  $X \sim \text{Bi}(n, \frac{1}{26})$ . We are looking for the value of  $n$  for which  $P(X \geq 5) = 0.95$ , i. e.,  $F_X(4) = P(X \leq 4) = 0.05$ . In order to find it, we use the de Moivre-Laplace theorem and approximate the binomial distribution with the normal distribution (we assume that  $n$  is large, so the approximation error will be small). By **F**, the expected value of  $X$  is  $E X = \frac{n}{26}$ , and its variance is  $\text{var } X = \frac{25n}{26^2}$ . Denoting the corresponding standardized variable by  $Z$ , we can reformulate the condition as

$$0.05 = P(X \leq 4) = P\left(Z \leq \frac{4 - \frac{n}{26}}{\frac{5\sqrt{n}}{26}}\right) = F_Z\left(\frac{104 - n}{5\sqrt{n}}\right),$$

where by the approximation assumption,  $F_Z \approx \Phi$  is the distribution function of the normal distribution  $N(0, 1)$ . Since we must have  $n > 104$ , using  $\Phi(-x) = 1 - \Phi(x)$ , the above equation gives  $n - 104 = \Phi^{-1}(0.95) \cdot 5\sqrt{n}$ . Using a table of the normal distribution or appropriate software, we can learn that  $z(0.95) = 1.65$ . Solving this quadratic equation, we get  $n \doteq 228.8$ . Thus, Vera must buy at least 229 yogurts.  $\square$

**10.H.4.** We roll a die 1200 times. Find the probability that the number of 6s lies between 150 and 250 (inclusive) using Chebyshev's inequality, and then using Moivre-Laplace theorem.

**Solution.** Let  $X$  denote the random variable which corresponds to the number of 6s. Clearly,  $X \sim \text{Bi}(1200, \frac{1}{6})$ . By **F**, we have  $E X = 1200 \cdot \frac{1}{6} = 200$  and  $\text{var } X = 200(1 - \frac{1}{6}) = \frac{500}{3}$ . The condition on the number of 6s says that  $150 \leq X \leq 250$ , which can be written as  $|X - 200| \leq 50$ . Using Chebyshev's inequality **10.2.32**, we get

$$P(|X - 200| \leq 50) = 1 - P(|X - 200| \geq 51) \geq 1 - \frac{500}{3 \cdot 51^2} \approx 0.94.$$

(2) The exact value of the wanted probability is given by the expression

$$P(150 \leq X \leq 250) = F_X(250) - F_X(150),$$

where  $F_X$  is the distribution function of the binomial distribution. By definition,

$$P(150 \leq X \leq 250) = \sum_{k=150}^{250} \binom{1200}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{1200-k}.$$

This expression is hard to evaluate without a computer, so we use Moivre-Laplace theorem. Replacing  $X$  with the standardized random variable

$$Z = \frac{\sqrt{3}(X - 200)}{10\sqrt{5}},$$

is the random variable corresponding to the won amount, it seems that the correct answer is "anything below the expected value  $E X$ ".

As derived in **10.2.1**,  $P(T = k) = 2^{-k}$ , provided that the coin is fair. Sum up all the probabilities multiplied by  $2^k$ , to obtain  $\sum_1^\infty 1 = \infty$ . Therefore, the expected value does not exist. So it seems that it is advantageous for the gambler to play even if the initial amount is very high...

Simulate the game for a while, to obtain that the amount won is somewhere around  $2^4$ . The reason is that no one is able to play infinitely long, hence the extremely high amounts are not feasible enough to be won, so such amounts cannot be taken seriously. In decision theory, these cases (when the expected value does not directly correspond to the evaluated utility) are called *St. Petersburg paradox*, and much literature has been devoted to this topic.<sup>3</sup>

**10.2.29. Properties of the expected value.** In the case of simple distributions, compute the expected value directly from the definition. For instance, for the Bernoulli distribution  $A(p)$ , it is immediate that

$$E X = (1 - p) \cdot 0 + p \cdot 1 = p.$$

Similarly, compute the expected value  $np$  of the binomial distribution  $\text{Bi}(n, p)$ . This requires more thought. The result is a direct corollary of the following general theorem since  $\text{Bi}(n, p)$  is the sum of  $n$  random variables with the Bernoulli distributions  $A(p)$ .

For any random variables  $X, Y$ , real constants  $a, b$ , consider the expected values of the functions of random variables  $X + Y$  and  $a + bX$ , provided the expected values  $E X$  and  $E Y$  exist.

It follows directly from the definition that the constant random variable  $a$  has expected value  $a$ . Further,

$$E(bX) = b E X,$$

since the constant  $b$  can be factored out from the sums or integrals.

More generally, the expected value of the product of independent random variables  $X$  and  $Y$  can be computed as follows. Suppose the components of the vector  $(X, Y)$  are discrete and independent, with probability mass functions  $f_X(x_i), f_Y(y_j)$ . Then,

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j f_X(x_i) f_Y(y_j) \\ &= \left(\sum_i x_i f_X(x_i)\right) \left(\sum_j y_j f_Y(y_j)\right) = E X E Y. \end{aligned}$$

Similarly, verify the equality  $E(XY) = E X E Y$  for independent continuous random variables.

<sup>3</sup>Going back to Bernoulli, 1738, the real value is given by the utility, rather than the price.

then, by 10.2.40, we have  $Z \sim N(0, 1)$ , i. e.,  $F_Z \approx \Phi$ . Thus,

$$P(150 \leq X \leq 250) = P\left(\frac{\sqrt{3}(250-200)}{10\sqrt{5}} \leq Z \leq \frac{\sqrt{3}(150-200)}{10\sqrt{5}}\right) \approx \Phi(\sqrt{15}) - \Phi(-\sqrt{15}) = 2\Phi(\sqrt{15}) - 1.$$

We learn that  $\Phi(\sqrt{15}) \approx 0.99994$ , so the wanted probability is approximately 99.988 %.  $\square$

**10.H.5.** At the Faculty of Informatics, 10 % of students have prumer less than 1.2 (let us call them successful). How many students must we meet if the probability that there are 8–12 % successful ones among them is to be at least 0.95? Solve this problem using Chebyshev’s inequality, and then using Moivre-Laplace theorem.

**Solution.** Let  $X$  denote the random variable that corresponds to the number of successful students, parametrized by the number  $n$  of students we meet. Since a randomly met student has probability 10 % of being successful, when meeting  $n$  students, we have  $X \sim \text{Bi}(n, \frac{1}{10})$ . By F, we have  $E X = 0.1n$  and  $\text{var } X = 0.09n$ . By Chebyshev’s inequality 10.2.32, the wanted probability satisfies

$$P(|X - 0.1n| \leq 0.02n) = 1 - P(|X - 0.1n| \geq 0.02n) \geq 1 - \frac{0.1 \cdot 0.9n}{(0.02n)^2} = 1 - \frac{225}{n}.$$

The inequality  $1 - \frac{225}{n} \geq 0.95$  and hence

$$P(|X - 0.1n| \leq 0.02n) \geq 0.95$$

holds for  $n \geq 4500$ . The exact value of the probability is given in terms of the distribution function  $F_X$  of the binomial distribution:

$$P(0.08n \leq X \leq 0.12n) = F_X(0.12n) - F_X(0.08n).$$

Using the de Moivre-Laplace theorem (see 10.2.40), we can approximate the standardized random variable  $Z = \frac{10X-n}{3\sqrt{n}}$  with the standard normal distribution,  $F_Z \approx \Phi$ , so

$$\begin{aligned} 0.95 &= P(0.08n \leq X \leq 0.12n) = P\left(-\frac{\sqrt{n}}{15} \leq Z \leq \frac{\sqrt{n}}{15}\right) \\ &\approx \Phi\left(\frac{\sqrt{n}}{15}\right) - \Phi\left(-\frac{\sqrt{n}}{15}\right) = \\ &= 2\Phi\left(\frac{\sqrt{n}}{15}\right) - 1. \end{aligned}$$

Hence  $\sqrt{n} = 15z(0.975)$  and we learn  $n \approx 864.4$ . Thus, we can see that it is sufficient to meet 865 students.  $\square$

**10.H.6.** The probability that a planted tree will grow is 0.8. What is the probability that out of 500 planted trees, at least 380 trees will grow?

**Solution.** The random variable  $X$  that corresponds to the number of trees that will grow has binomial distribution  $X \sim \text{Bi}(500, \frac{4}{5})$ . By F, we have  $E X = 400$  and  $\text{var } X = 80$ . The standardized random variable is  $Z = \frac{X-400}{\sqrt{80}}$ . By the de

Now compute  $E(X + Y)$  for arbitrary random variables. For discrete distributions of  $X$  and  $Y$ ,

$$\begin{aligned} E(X + Y) &= \sum_i \sum_j (x_i + y_j) P(X = x_i, Y = y_j) \\ &= \sum_i \left( x_i \sum_j P(X = x_i, Y = y_j) \right) \\ &\quad + \sum_j \left( y_j \sum_i P(X = x_i, Y = y_j) \right) \\ &= \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j), \end{aligned}$$

where absolute convergence of the first double sum follows from the triangle inequality and the absolute convergence of the sums that stand for the expected values of the particular random variables. Absolute convergence is used in order to interchange the sums.

Dealing with continuous variables  $X$  and  $Y$ , whose expected values exist, proceed analogously.

$$\begin{aligned} E(X + Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx + \int_{-\infty}^{\infty} y f_Y(y) dy = E X + E Y, \end{aligned}$$

where absolute convergence of integrals of the expected values  $E X$  and  $E Y$  is used to interchange the integrals by Fubini’s theorem.

Altogether, the expected formula:

$$E(X + Y) = E X + E Y$$

is obtained, whenever the expected values  $E X$  and  $E Y$  exist.

Straightforward application of this result leads to the following:

AFFINE NATURE OF EXPECTED VALUES

For any constants  $a, b_1, \dots, b_k$  and random variables  $X_1, \dots, X_k$ ,

$$E(a + b_1 X_1 + \dots + b_k X_k) = a + b_1 E X_1 + \dots + b_k E X_k.$$

The following theorem extends this behaviour with respect to affine transformations of random vectors, and shows that the expected value is invariant with respect to affine transformations, as is the arithmetic mean:

**Theorem.** Let  $X = (X_1, \dots, X_n)$  be a random vector with expected value  $E X$ ,  $a \in \mathbb{R}^m$ ,  $B \in \text{Mat}_{mn}(\mathbb{R})$  a matrix. Then,

$$E(a + B \cdot X) = a + B \cdot E X.$$

**PROOF.** There is almost nothing remaining to be proved. Since the expected value of a vector is defined as the vector of the expected values of the components, it suffices to restrict attention to a single item in  $E(a + B \cdot X)$ . Thus, it can be assumed that  $a$  is a scalar and  $B$  is a matrix with a single row.

Moivre-Laplace theorem, we have  $F_Z \approx \Phi$ , so

$$\begin{aligned} P(X \geq 380) &= P\left(Z \geq \frac{380 - 400}{\sqrt{80}}\right) \approx 1 - \Phi\left(-\frac{\sqrt{20}}{2}\right) = \\ &= \Phi\left(\frac{\sqrt{20}}{2}\right) \approx 0.987. \end{aligned}$$

□

**10.H.7.** Using the distribution function of the standard normal distribution, find the probability that the absolute difference between the heads and the tails in 1600 tosses of a coin is at least 82.

**Solution.** Let  $X$  denote the random variable that corresponds to the number of times the coin came up heads. Then  $X$  has binomial distribution  $Bi(1600, 1/2)$  (with expected value 800 and standard deviation 20), so for a large value of  $n = 1600$ , by the de Moivre-Laplace theorem, the distribution function of the variable  $\frac{X-800}{20}$  can be approximated with the distribution function  $\Phi$  of the standard normal distribution. Thus, the wanted probability is

$$\begin{aligned} P &= 1 - P[759 \leq X \leq 841] \\ &= 1 - P\left[-2.05 \leq \frac{X - 800}{20} \leq 2.05\right] \\ &\doteq 2\Phi(-2, 05) \doteq 0.0404. \end{aligned}$$

□

**10.H.8.** Using the distribution function of the standard normal distribution, find the probability that the absolute difference between the heads and the tails in 3600 tosses of a coin is at most 66.

**Solution.** Let  $X$  denote the random variable that corresponds to the number of times the coin came up heads. Then  $X$  has binomial distribution  $Bi(3600, 1/2)$  (with expected value 1800 and standard deviation 30), so for a large value of  $n = 3600$ , the distribution function of the variable  $\frac{X-1800}{30}$  can be approximated, by the de Moivre-Laplace theorem, with the distribution function  $\Phi$  of the standard normal distribution. Thus, the wanted probability is

$$\begin{aligned} P[1767 \leq X \leq 1833] &= P\left[-1, 1 \leq \frac{X - 1800}{30} \leq 1, 1\right] \\ &\doteq \Phi(1, 1) - \Phi(-1, 1) \doteq 0, 7498. \end{aligned}$$

□

**10.H.9.** The probability that a seed will grow is 0.9. How many seeds must we plant if we require that with probability at least 0.995, the relative number of grown items differs from 0.9 by at most 0.034.

**Solution.** The random variable  $X$  that corresponds to the number of grown seeds, out of  $n$  planted ones, has binomial distribution  $X \sim Bi(n, \frac{9}{10})$ . By F, we have  $E X = 0.9n$  and  $\text{var } X = 0.09n$ , so the standardized variable is  $Z = \frac{X-0.9n}{\sqrt{0.09n}}$ .

Then, the expected value of a finite sum of random variables is obtained, and by the above results, that exists and is given as the sum of the expected values of the individual items. This is exactly what is wanted to be proved. □

**10.2.30. Quantiles and critical values.** Introduce numerical characteristics that are analogous to those from descriptive statistics. There, the next useful characteristics are the *quantiles*, cf. 10.1.5.



Consider a random variable  $X$  whose distribution function  $F_X$  is strictly monotone. This is satisfied by any random variable whose density is nowhere equal to zero, which is the case for the normal distribution, for example. In this case, define the *quantile function*  $F_X^{-1}$  simply as the inverse function  $(F_X)^{-1} : (0, 1) \rightarrow \mathbb{R}$ . This means that the value  $y = F_X^{-1}(\alpha)$  is such that  $P(X < y) = \alpha$ . This corresponds precisely to the quantiles from descriptive statistics using relative frequencies for the probabilities.

#### QUANTILE FUNCTION

For any random variable  $X$  with distribution function  $F_X(x)$ , define its *quantile function*

$$F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R}; F(x) \geq \alpha\}, \alpha \in (0, 1).$$

Clearly, this is a generalization of the previous definition in the case the distribution function is strictly monotone.

As seen in descriptive statistics, the most used quantiles are for  $\alpha = 0.5$  (the *median*),  $\alpha = 0.25$  (the *first quartile*),  $\alpha = 0.75$  (the *third quartile*). Similarly for *deciles* and *percentiles* when  $\alpha$  is equal to (integer) multiples of tenths and hundredths, respectively.

It follows directly from the definition that the quantile function for a given random variable  $X$  allows the determination of intervals into which the values of  $X$  fall with a chosen probability. For instance, the value  $F_X^{-1}(0.975)$ , approximately 1.96, corresponds to percentile 97.5 for the normal distribution  $N(0, 1)$ . This says that with the probability of 2.5%, the value of such a random variable  $Z \sim N(0, 1)$  is at least 1.96. Since the density of the variable  $Z$  is symmetric with respect to the origin, this observation can be interpreted as that there is only a 5% probability that the value of  $|Z|$  is greater 1.96.

There are similar intervals and values when discussing the reliability of estimates of characteristics of random variables.

#### CRITICAL VALUES

For a random variable  $X$  and a real number  $0 < \alpha < 1$ , define its *critical value*  $x(\alpha)$  at level  $\alpha$  as

$$P(X \geq x(\alpha)) = \alpha.$$

This means that  $x(\alpha) = F_X^{-1}(1 - \alpha)$  where  $F_X^{-1}$  is the quantile function of the random variable  $X$ .

The condition in question can be written as

$$\begin{aligned} P(|X - 0.9n| \leq 0.034n) &= P\left(|Z| \leq \frac{0.034n}{\sqrt{0.09n}}\right) = \\ &= P\left(|Z| \leq \frac{0.34}{3}\sqrt{n}\right) \geq 0.995. \end{aligned}$$

By the de Moivre-Laplace theorem, for large  $n$ , the distribution function can be approximated by the distribution function  $\Phi$  of the normal distribution. Thus,

$$\begin{aligned} P\left(|Z| \leq \frac{0.34}{3}\sqrt{n}\right) &\approx \Phi\left(\frac{0.34}{3}\sqrt{n}\right) - \Phi\left(-\frac{0.34}{3}\sqrt{n}\right) = \\ &= 2\Phi\left(\frac{0.34}{3}\sqrt{n}\right) - 1. \end{aligned}$$

Altogether, we get the condition

$$2\Phi\left(\frac{0.34}{3}\sqrt{n}\right) - 1 \geq 0.995.$$

Odtud vypočítáme  $n \geq \left(\frac{3z(0.9975)}{0.34}\right)^2 \approx 615$ . □

**10.H.10.** The service life (in hours) of a certain kind of gadget has exponential distribution with parameter  $\lambda = \frac{1}{10}$ . Using the central limit theorem, bound the probability that the total service life of 100 such gadgets lies between 900 and 1050 hours.

**Solution.** In exercise 10.F.5, we computed that the expected value and variance of a random variable  $X_i$  with exponential distribution are equal to  $E X_i = \frac{1}{\lambda}$  and  $\text{var } X_i = \frac{1}{\lambda^2}$ , respectively. Thus, the expected service life of each gadget is  $E X_i = \mu = 10$  hours, with variance  $\text{var } X_i = \sigma^2 = 100$  hours<sup>2</sup>. By the central limit theorem, the distribution of the transformed random variable  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{100} \sum_{i=1}^{100} X_i - 10$  approaches the standard normal distribution as  $n$  tends to infinity. Thus, the wanted probability for the service life of 100 gadgets

$$P(900 \leq \sum X_i \leq 1050) = P\left(-1 \leq \frac{1}{100} \sum_{i=1}^{100} X_i - 10 \leq 0.5\right)$$

can be approximated with the distribution function of the normal distribution:

$$P(900 \leq \sum X_i \leq 1050) \approx \Phi(0.5) - \Phi(-1) \approx 0.533.$$

□

**10.H.11.** We keep putting items into a chest. The expected mass of an item is 3 kg and the standard deviation is 0.8 kg. What is the maximum number of items that we can put into the chest so that with probability at least 99%, the total mass does not exceed one ton?

**Solution.** Let  $X_i$  denote the random variable that corresponds to the mass of the  $i$ -th item. Then, we have  $\mu = E X_i = 3$  and  $\sigma = \sqrt{\text{var } X_i} = 0.8$  (in kilograms), and we want to have

$$P\left(\sum_{i=1}^n X_i \leq 1000\right) = 0.99.$$

**10.2.31. Variance and standard deviation.** The simple numerical characteristics concerning the variability of sample values in descriptive statistics were the variance and the standard deviation. Define them similarly for random variables.



VARIANCE OF A RANDOM VARIABLE

Given a random variable  $X$  with finite expected value, its *variance* is defined as

$$\text{var } X = E((X - E X)^2),$$

provided the right-hand expected value exists. Otherwise, the variance of  $X$  does not exist.

The square root  $\sqrt{\text{var } X}$  of the variance is called the *standard deviation of the random variable*  $X$ .

Using the properties of the expected value, a simpler formula can be derived for the variance of a random variable  $X$  whose expected value exists:

$$\begin{aligned} \text{var } X &= E(X - E X)^2 = E(X^2 - 2X(E X) + (E X)^2) \\ &= E X^2 - 2(E X)^2 + (E X)^2 \\ &= E X^2 - (E X)^2. \end{aligned}$$

Consider how affine transformations change the variance of a random variable. Given real numbers  $a, b$  and a random variable  $X$  with expected value and variance, consider the random variable  $Y = a + bX$ . Compute

$$\begin{aligned} \text{var } Y &= E((a + bX) - E(a + bX))^2 = E(b(X - E X))^2 \\ &= b^2 \text{var } X. \end{aligned}$$

Thus are derived the following useful formulae:

PROPERTIES OF VARIANCE

$$\begin{aligned} (1) \quad &\text{var } X = E(X^2) - (E X)^2 \\ (2) \quad &\text{var}(a + bX) = b^2 \text{var } X \\ (3) \quad &\sqrt{\text{var}(a + bX)} = b\sqrt{\text{var } X} \end{aligned}$$

Given a random variable  $X$  with expected value and non-zero variance, define its *standardization* as the random variable

$$Z = \frac{X - E X}{\sqrt{\text{var } X}}.$$

Thus, the standardized variable is the affine transformation of the original variable whose expected value equals zero and variance equals one.

**10.2.32. Chebyshev's inequality.** A good illustration of the usefulness of variance is the Chebyshev's inequality. This connects the variance directly to the probability that the random variable assumes values that are distant from its expected value.



By the central limit theorem 10.2.40, the distribution of the random variable

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{X_i - 3}{0.8} \right) = \frac{1}{0.8\sqrt{n}} \sum_{i=1}^n X_i - \frac{3\sqrt{n}}{0.8}$$

can be approximated by the standard normal distribution. Thus, we get

$$P\left(\sum_{i=1}^n X_i \leq 1000\right) = P\left(S_n \leq \frac{1000}{0.8\sqrt{n}} - \frac{3\sqrt{n}}{0.8}\right) \approx \Phi\left(\frac{1000}{0.8\sqrt{n}} - \frac{3\sqrt{n}}{0.8}\right)$$

We learn that  $z(0.99) \approx 2.326$ , so the wanted  $n$  satisfies the quadratic equation

$$\frac{1000}{0.8\sqrt{n}} - \frac{3\sqrt{n}}{0.8} = 2.326,$$

whence we get  $n \approx 322$ .  $\square$

### I. Testing samples from the normal distribution

In subsection 10.3.4, we introduced the so-called two-sided interval estimate of an unknown parameter  $\mu$  of the normal distribution  $N(\mu, \sigma^2)$ . In some cases, we may be interested only in an upper or lower estimate, i.e. a statistic  $U$  or  $L$  for which  $P(\mu < U)$  or  $P(L < \mu)$ , respectively. Then, we talk about a one-sided confidence interval  $(-\infty, U)$  or  $(L, \infty)$ . The formula for these intervals can be derived similarly as for the two-sided interval. Now, we have for the random variable  $Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$  that

$$1 - \alpha = \Phi(z(1 - \alpha)) = P(Z < z(1 - \alpha)).$$

Hence it immediately follows that

$$1 - \alpha = P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z(1 - \alpha) < \mu\right),$$

so  $L = \bar{X} - \frac{\sigma}{\sqrt{n}} z(1 - \alpha)$ . Similarly, we find  $U = \bar{X} + \frac{\sigma}{\sqrt{n}} z(1 - \alpha)$ , and for a distribution with unknown variance,  $\mu \geq \bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1 - \alpha)$  and  $\mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(1 - \alpha)$ .

If we want to estimate the variance  $\sigma^2$  of a random distribution, then we use theorem 10.3.3, similarly as when we derived it for the expected value. This time, we use the second part of the theorem, by which the random variable  $\frac{n-1}{\sigma^2} S^2$  has distribution  $\chi^2$ . Then, we can immediately see that

$$1 - \alpha = P\left(\chi_{n-1}^2(\alpha/2) \leq \frac{n-1}{\sigma^2} S^2 \leq \chi_{n-1}^2(1 - \alpha/2)\right).$$

Thus, the two-sided  $100(1 - \alpha)\%$  confidence interval for the variance is

$$\left( \frac{(n-1)S^2}{\chi_{n-1}^2(1 - \alpha/2)}, \frac{(n-1)S^2}{\chi_{n-1}^2(\alpha/2)} \right)$$

and similarly for the one-sided upper and lower estimates, we get

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi_{n-1}^2(\alpha)}, \text{ resp. } \frac{(n-1)S^2}{\chi_{n-1}^2(1 - \alpha)} \leq \sigma^2.$$

### CHEBYSHEV'S INEQUALITY

**Theorem.** Consider a random variable  $X$  with finite variance, and fix an arbitrary  $\varepsilon > 0$ . Then,

$$P(|X - E X| \geq \varepsilon) \leq \frac{\text{var } X}{\varepsilon^2}.$$

**PROOF.** Suppose  $X$  is continuous. Set  $\mu = E X$  and compute, using the definition:

$$\begin{aligned} \text{var } X &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ &= \int_{|x-\mu| \geq \varepsilon} (x - \mu)^2 f(x) dx \\ &\quad + \int_{|x-\mu| < \varepsilon} (x - \mu)^2 f(x) dx \\ &\geq \int_{|x-\mu| \geq \varepsilon} \varepsilon^2 f(x) dx = \varepsilon^2 P(|X - \mu| \geq \varepsilon). \end{aligned}$$

$\square$

The analogous proof for discrete random variables is left as an exercise for the reader.

Realizing that the variance is the square of the standard deviation  $\sigma$ , the choice  $\varepsilon = k\sigma$  yields the probability

$$P(|X - E X| \geq k\sigma) \leq \frac{1}{k^2}.$$

The Chebyshev's inequality helps understanding asymptotic descriptions of limit processes. For instance, consider the sequence of random variables  $X_1, X_2, \dots$  with probability distributions  $X_n \sim \text{Bi}(n, p)$ , with a fixed value of  $p$ ,  $0 < p < 1$ . Intuitively, it is expected that the relative frequency of success should approach the probability  $p$  as  $n$  increases, i.e., that the values of the random variables  $Y_n = \frac{1}{n} X_n$  should approach  $p$ . Clearly,

$$E Y_n = \frac{np}{n} = p, \quad \text{var } Y_n = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Direct application of Chebyshev's inequality yields, for any fixed  $\varepsilon > 0$ , that

$$P(|Y_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}.$$

Hence it is clear that, for any fixed  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) = 0.$$

This result is known as *Bernoulli's theorem* (one of many).

This type of limit behaviour is called *convergence in probability*. Thus it is proved (as a corollary of Chebyshev's inequality) that the random variables  $Y_n$  converge in probability to the constant random variable  $p$ .

**10.I.1.** We roll a die 600 times, obtaining only 45 sixes. Is it possible to say that the die is ideal at level  $\alpha = 0.01$ ?

**Solution.** For an ideal die, the probability of rolling a six is always  $p = \frac{1}{6}$ . The number of sixes in 600 rolls is given by a random variable  $X$  with binomial distribution  $X \sim \text{Bi}(600, \frac{1}{6})$ . By 10.2.40, this distribution can be approximated by the distribution  $N(100, \frac{250}{3})$ . The measured value  $X = 45$  can be considered a random sample consisting of one item. Assuming that the variance is known and applying 10.3.4, we get that the 99% (two-sided) confidence interval for the expected value  $\mu$  equals  $(45 - \sqrt{\frac{250}{3}}z(0.995), 45 + \sqrt{\frac{250}{3}}z(0.995))$ . We learn that the quantile is approximately  $z(0.995) \approx 2.58$ , which gives the interval (21, 69). However, for an ideal die, we clearly have  $\mu = 100$ , so our die is not ideal at level  $\alpha = 0.01$ .  $\square$

**10.I.2.** Suppose the height of 10-years-old boys has normal distribution  $N(\mu, \sigma^2)$  with unknown expected value  $\mu$  and variance  $\sigma^2 = 39.112$ . Taking the height of 15 boys, we get the sample mean  $\bar{X} = 139.13$ . Find

- i) the 99% two-sided confidence interval for the parameter  $\mu$ ,
- ii) the lower estimate for  $\mu$  at significance level 95 %.

**Solution.** a) By 10.3.4, the  $100(1 - \alpha)\%$  two-sided confidence interval for the unknown expected value  $\mu$  of the normal distribution is

$$(1) \quad \mu \in \left( \bar{X} - \frac{\sigma}{\sqrt{n}}z(1 - \alpha/2), \bar{X} + \frac{\sigma}{\sqrt{n}}z(1 - \alpha/2) \right),$$

where  $\bar{X}$  is the sample mean of  $n$  items,  $\sigma^2$  is the known variance, and  $z(1 - \alpha/2)$  is the corresponding quantile. Substituting the given values  $n = 15$ ,  $\sigma \approx 6.254$  and the learned  $z(0.995) \approx 2.576$ , we get  $\frac{\sigma}{\sqrt{n}}z(\alpha/2) \approx 4.16$ , i. e.,  $\mu \in (134.97, 143.29)$ .

b) The lower estimate  $L$  for the parameter  $\mu$  at significance level 95 % is given by the expression  $L = \bar{X} - \frac{\sigma}{\sqrt{n}}z(0.95)$ . We learn that  $z(0.95) \approx 1.645$ , and direct substitution leads to  $\mu \in (136.474, \infty)$ .  $\square$

**10.I.3.** A customer tests the quality of bought products by examining 21 randomly chosen ones. He will accept the delivery if the sample standard deviation does not exceed 0.2 mm. We know that the pursued property of the products has normal distribution of the form  $N(10 \text{ mm}; 0.0734 \text{ mm}^2)$ . Using statistical tables, find the probability that the delivery will be accepted. How does the answer change if the customer, in order to save expenses, tests only 4 products?

**Solution.** The problem asks for the probability  $P(S \leq 0.2)$ . By theorem 10.3.3, when sampling  $n$  products, the random variable  $\frac{n-1}{\sigma^2}S^2$  has distribution  $\chi_{n-1}^2$ . In our case,  $n = 21$  and  $\sigma^2 = 0.0734$ , so

$$P(S \leq 0.2) = P\left(\frac{20}{0.0734}S^2 \leq \frac{20}{0.0734}0.2^2\right) = \chi_{20}^2\left(\frac{20 \cdot 0.2^2}{0.0734}\right) = bdE((X - E X)(Y - E Y)) = bd \text{cov}(X, Y).$$

**10.2.33. Covariance.** We return to random vectors. In the case of the expected value, the situation is very simple — just take the vector of expected values. When characterizing the variability, the dependencies between the individual components are also of much interest. We follow the idea from 10.1.9 again.



COVARIANCE

Given random variables  $X, Y$  whose variances exist, Define their covariance as

$$\text{cov}(X, Y) = E((X - E X)(Y - E Y))$$

The basic properties of the concept can be derived very easily:

**Theorem.** For any random variables  $X, Y, Z$  whose variances exist and real numbers  $a, b, c, d$ ,

- (1)  $\text{cov}(X, Y) = \text{cov}(Y, X)$
- (2)  $\text{cov}(X, Y) = E(XY) - (E X)(E Y)$
- (3)  $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$
- (4)  $\text{cov}(a + bX, c + dY) = bd \text{cov}(X, Y)$
- (5)  $\text{var}(X + Y) = \text{var} X + \text{var} Y + 2 \text{cov}(X, Y)$ .

Moreover, if  $X$  and  $Y$  are independent, then  $\text{cov}(X, Y) = 0$ , and consequently

$$(6) \quad \text{var}(X + Y) = \text{var} X + \text{var} Y.$$

**PROOF.** Directly from the definition, the covariance is symmetric in the arguments. The second proposition follows immediately from the properties of the expected value:

$$\begin{aligned} \text{cov}(X, Y) &= E(X - E X)(Y - E Y) \\ &= E(XY) - (E Y)X - (E X)Y + E X E Y \\ &= E(XY) - E X E Y. \end{aligned}$$

The next proposition also follows easily if the definition is expanded and the fact that the expected value of the sum of random variables equals the sum of their expected values is used.

The next proposition can be computed directly:

$$\begin{aligned} \text{cov}(a + bX, c + dY) &= \\ &= E((a + bX - E(a + bX))(c + dY - E(c + dY))) \\ &= E((bX - bE(X))(dY - dE(Y))) \\ &= E(bd(X - E(X))(Y - E(Y))) \\ &= bdE((X - E X)(Y - E Y)) = bd \text{cov}(X, Y). \end{aligned}$$

The expression in the argument of the distribution function is approximately 10.9, and we can learn from the table of the  $\chi^2$  distribution that  $\chi_{20}^2(10.9) \approx 0.05$ . Thus, the probability that delivery will be accepted is only 5%. We could have expected the probability to be low: indeed,  $ES^2 = \sigma^2 = 0.0734 > 0.2^2$ . If the customer tests only 4 products, then the probability of acceptance is given by the expression  $\chi_3^2\left(\frac{3 \cdot 0.2^2}{0.0734}\right) \approx \chi_3^2(1.63)$ . The value of the distribution function of  $\chi^2$  in this argument cannot be found in most tables. Therefore, we estimate it using linear interpolation. For instance, if the nearest known points are  $\chi_3^2(0.58) = 0.1$  and  $\chi_3^2(6.25) = 0.9$ , then

$$\chi_3^2(1.63) \approx (1.63 - 0.58) \frac{0.9 - 0.1}{6.25 - 0.58} + 0.1 \approx 0.24.$$

Although this results is only an estimate, we can be sure that the probability of acceptance is much greater than when testing 21 products.  $\square$

**10.I.4.** From a population with distribution  $N(\mu, \sigma^2)$ , where  $\sigma^2 = 0.06$ , we have sampled the values 1.3; 1.8; 1.4; 1.2; 0.9; 1.5; 1.7. Find the two-sided 95% confidence interval for the unknown expected value.

**Solution.** We have a random sample of size  $n = 7$  from the normal distribution with known variance  $\sigma^2 = 0.06$ . The sample mean is

$$\bar{X} = \frac{1}{7}(1.3 + 1.8 + 1.4 + 1.2 + 0.9 + 1.5 + 1.7) = 1.4$$

and we can learn for the given confidence level  $\alpha = 0.05$  that  $z(1 - \alpha/2) = z(0.975) \approx 1.96$ . Substituting into (1), we immediately obtain the wanted interval (1.22, 1.58).  $\square$

**10.I.5.** Let  $X_1, \dots, X_n$  be a random sample from the distribution  $N(\mu, 0.04)$ . Find the least number of measurements that are necessary so that the length of the 95% confidence interval for  $\mu$  would not exceed 0.16.

**Solution.** Since we have a normal distribution with known variance, we know from (1) that the length of the  $(1 - \alpha)\%$  confidence interval is  $\frac{2\sigma}{\sqrt{n}}z(1 - \alpha/2)$ . Substituting the given values, we get that the number  $n$  of measurements satisfies the inequality

$$\frac{2 \cdot 0.2}{\sqrt{n}}z(0.975) \leq 0.16.$$

Since  $z(0.975) \approx 1.96$ , we obtain  $n \geq 24.01$ . Thus, at least 25 measurements are necessary.  $\square$

**10.I.6.** Consider a random variable  $X$  with distribution  $N(\mu, \sigma^2)$ , where  $\mu, \sigma^2$  are unknown. The following table shows the frequencies of individual values of this random variable:

$X_i$	8	11	12	14	15	16	17	18	20	21
$n_i$	1	2	3	4	7	5	4	3	2	1

Calculate the sample mean, sample variance, sample standard deviation, and find the 99% confidence interval for the expected value  $\mu$ .

The other propositions about the variance are quite simple corollaries:

$$\begin{aligned} \text{var}(X + Y) &= E((X + Y) - E(X + Y))^2 \\ &= E((X - EX) + (Y - EY))^2 \\ &= E(X - EX)^2 + 2E(X - EX)(Y - EY) \\ &\quad + E(Y - EY)^2 \\ &= \text{var } X + 2\text{cov}(X, Y) + \text{var } Y. \end{aligned}$$

Furthermore, if  $X$  and  $Y$  are independent, then  $E(XY) = EXEY$ , and hence that their covariance is zero.  $\square$

Directly from the definition,

$$\text{var}(X) = \text{cov}(X, X).$$

The latter theorem claims that covariance is a symmetric bilinear form on the real vector space of random variables whose variance exists. The variance is the corresponding quadratic form. The covariance can be computed from the variance of the particular random variables and of their sum, as seen in linear algebra, see the property (5).

Notice that the random variable, equal to the sum of  $n$  independent and identically distributed random variables  $Y_i$  behaves, very much differently than the multiple  $nY$ . In fact,

$$\text{var}(Y_1 + \dots + Y_n) = n \text{var } Y, \quad \text{var}(nY) = n^2 \text{var } Y.$$

**10.2.34. Correlation of random variables.** To a certain extent, covariance corresponds to dependency between the random variables. Its relative version is called the *correlation of random variables* and, similarly as for the standard deviation, the following concept is defined:



#### CORRELATION COEFFICIENT

The *correlation coefficient* of random variables  $X$  and  $Y$  whose variances are finite and non-zero is defined as

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}}.$$

As seen from theorem 10.2.33, the correlation coefficient of random variables equals the covariance of the standardized variables  $\frac{1}{\sqrt{\text{var } X}}(X - EX)$  and  $\frac{1}{\sqrt{\text{var } Y}}(Y - EY)$ .

The following equalities hold (here,  $a, b, c, d$  are real constants,  $bd \neq 0$ , and  $X, Y$  are random variables with finite non-zero variances)

$$\begin{aligned} \rho_{a+bX, c+dY} &= \text{sgn}(bd)\rho_{X,Y} \\ \rho_{X, X} &= 1. \end{aligned}$$

Moreover, if  $X$  and  $Y$  are independent, then  $\rho_{X,Y} = 0$ .

Note that if the variance of a random variable  $X$  is zero, then it assumes the value  $EX$  with probability 1. If the value of  $X$  falls into an interval  $I$  not containing  $EX$  with probability  $p \neq 0$ , then the expression  $\text{var } X = E(X - EX)^2$  is positive. Stochastically, random variables with zero variance behave as constants.

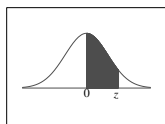
**Solution.** The sample mean is given by the expression  $\bar{X} = \sum n_i X_i / \sum n_i$ . Substituting the given values, we get  $\bar{X} = 490/32 \approx 15.3$ . By definition, the sample variance is  $S = \sum n_i (X_i - \bar{X})^2 / (\sum n_i - 1)$ . Substituting the given values, we get  $S^2 = 1943/256 \approx 7.6$ , so the sample standard deviation is  $S \approx 2.8$ . The formula for the two-sided  $(1 - \alpha)\%$  confidence interval for the expected value  $\mu$ , when the variance is unknown, was derived at the end of subsection 10.3.4:

$$\mu \in \left( \bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(1 - \alpha/2), \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(1 - \alpha/2) \right).$$

Substitution yields  $\bar{X} = 15.3$ ,  $n = 32$ ,  $S \approx 2.8$ ,  $\alpha = 0.01$ , and we learn  $t_{31}(0.995) \approx 2.75$ . Thus, the 99% confidence interval is  $\mu \in (14.0, 16.7)$ .  $\square$

**10.I.7.** Using the following table of the distribution function of the normal distribution, find the probability that the absolute difference between the heads and the tails in 3600 tosses of a coin is greater than 90.

Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998

Gilbert Strang, Typset with JTBK on April 20, 2006.

**Solution.** Let  $X$  denote the random variable that corresponds to the number of heads. Then,  $X$  has binomial distribution  $Bi(3600, 1/2)$  (with expected value 1800 and standard deviation 30), so by the de Moivre-Laplace theorem, for large values of  $n$ , the distribution function of the variable  $\frac{X-1800}{30}$  can be approximated by the distribution function  $\Phi$  of the normal

If the covariance is a positive-definite symmetric bilinear form, then it would follow from the Cauchy-Schwarz inequality (see 3.4.3) that

$$(1) \quad |\rho_{X,Y}| \leq 1$$

The following theorem claims more. It shows that the full correlation or anti-correlation, i.e.  $\rho_{X,Y} = \pm 1$  of random variables  $X$  and  $Y$  says that they are bound by an affine relation  $Y = kX + c$ , where the sign of  $k$  corresponds to the sign in  $\rho_{X,Y} = \pm 1$ . On the other hand, a zero correlation coefficient says that the (potential) dependency between the variables is very far from any affine relation of the mentioned type. (Note, however, this does not mean that the variables must be independent).

For instance, consider random variables  $Z \sim N(0, 1)$  and  $Z^2$ . Then  $\text{cov}(Z, Z^2) = E Z^3 = 0$  since the density of  $Z$  is an even function. Thus the expected value of an odd power of  $Z$  is zero, if it exists.

**Theorem.** If the correlation coefficient is defined, then  $|\rho_{X,Y}| \leq 1$ . Equality holds if and only if there are constants  $k, c$  such that  $P(Y = kX + c) = 1$ .

**PROOF.** A stochastic affine relation between  $Y$  and  $X$  with nonzero coefficient at  $Y$  is sought. This is equivalent to  $Y + sX \sim D(c)$  for some fixed value of the parameter  $s$  and constant  $c$ . In such a case the variance vanishes. Thus one considers the following non-negative quadratic expression:

$$0 \leq \text{var} \left( \frac{Y - EY}{\sqrt{\text{var}Y}} + t \frac{X - EX}{\sqrt{\text{var}X}} \right) = 1 + 2t\rho_{X,Y} + t^2.$$

The right-hand quadratic expression does not have two distinct real roots; hence its discriminant cannot be positive. So  $4(\rho_{X,Y})^2 - 4 \leq 0$ . Hence the desired inequality is obtained, and also the discriminant vanishes if  $\rho_{X,Y} = \pm 1$ . For the only (double) root  $t_0$ , the corresponding random variable has zero variance; thus it assumes a fixed value with probability 1. This yields the affine relation as expected.  $\square$

**10.2.35. Covariance matrix.** The variability of a random vector must be considered. This suggests considering the covariances of all pairs of components. The following definition and theorem show that this leads to an analogy of the variance for vectors, including the behaviour of the variance under affine transformations of the random variables.





distribution. Thus, the wanted probability is

$$P = 1 - P[1755 \leq X \leq 1845] = 1 - P\left[-1.5 \leq \frac{X - 1800}{30} \leq 1.5\right] = 2\Phi(-1.5) \doteq 0.1336,$$

where the last value was learned from the table.  $\square$

**10.I.8.** The probability that a newborn baby is a boy is 0.515. Find the probability that there are at least the same number of girls as boys among ten thousand babies.

**Solution.**

$$P[X < 5000] = P\left[\underbrace{\frac{X - 5150}{\sqrt{5150 \cdot 0.485}}}_{\sim N(0,1)} < \underbrace{\frac{-150}{\sqrt{5150 \cdot 0.485}}}_{-3.001\dots}\right] \doteq 0.00135$$

$\square$

**10.I.9.** Using the distribution function of the standard normal distribution, find the probability that we get at least 3100 sixes out of 18000 rolls of a six-sided die.

**Solution.** We proceed similarly as in the exercises above.  $X$  has binomial distribution  $Bi(18000, 1/6)$ . We find the expected value  $((1/6)(18000) = 3000)$  as well as the standard deviation  $\sqrt{((1/6)(1 - 1/6)18000)} = 50$ . Therefore, the variable  $\frac{X-3000}{50}$  can be approximated with the distribution function  $\Phi$  of the standard normal distribution:

$$P[X \geq 3100] = P\left[\frac{X - 3000}{50} \geq \frac{3100 - 3000}{50}\right] = P\left[\frac{X - 3000}{50} \geq 2\right] \doteq 1 - \Phi(2) \doteq 0.0228.$$

$\square$

**10.I.10.** A public opinion agency organizes a survey of preferences of five political parties. How many randomly selected respondents must answer so that the probability that for each party, the survey result differs from the actual preference by no more than 2% is at least 0.95?

**Solution.** Let  $p_i, i = 1 \dots 5$  be the actual relative frequency of voters of the  $i$ -th political party in the population, and let  $X_i$  denote the number of voters of this party among  $n$  randomly chosen people. Note that given any five intervals, the events corresponding to  $X_i/n$  falling into the corresponding interval may be dependent. If we choose  $n$  so that for each  $i, X_i/n$  falls into the given interval with probability at least  $1 - ((1 - 0.95)/5) = 0.99$ , then the desired condition is sure to hold even in spite of the dependencies. Thus, let us look for  $n$  such that  $P[|\frac{X}{n} - p| < 0.02] \geq 0.99$ . First of all, we

COVARIANCE MATRIX

Consider a random vector  $X = (X_1, \dots, X_n)^T$  all of whose components have finite variances.

The *covariance matrix* of the random vector  $X$  is defined in terms of the expected value as (notice the vector  $X$  is viewed as a column of random variables now)

$$\text{var } X = E(X - E X)(X - E X)^T.$$

Using the definition of the expected value of a vector and expanding the matrix multiplication, it is immediate that the covariance matrix  $\text{var } X$  is the symmetric matrix

$$\begin{pmatrix} \text{var } X_1 & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var } X_2 & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var } X_n \end{pmatrix}.$$

**Theorem.** Consider a random vector  $X = (X_1, \dots, X_n)^T$  all of whose components have finite variances. Further, consider the transformed random vector  $Y = BX + c$ , where  $B$  is an  $m$ -by- $n$  matrix of real constants and  $c \in \mathbb{R}^m$  is a vector of constants. Then,

$$\text{var}(Y) = \text{var}(BX + c) = B(\text{var } X)B^T.$$

**PROOF.** The claim follows from direct computation, using the properties of the expected value:

$$\begin{aligned} \text{var}(Y) &= E((BX + c) - E(BX + c))((BX + c) - E(BX + c))^T \\ &= E(B(X - E X))(B(X - E X))^T \\ &= B E(X - E X)(X - E X)^T B^T \\ &= B(\text{var } X)B^T. \end{aligned}$$

$\square$

The constant part of the transformation has no impact, while with respect to the linear part of the transformation, the covariance matrix behaves as the matrix of a quadratic form.

**10.2.36. Moments and moment function.** The expected value and variance reflect the square of the deviation of values of a random variable from the average. In descriptive statistics, one also examines the skewness of the data, and it is natural to examine the variability of random variables in terms of higher powers of the given random variable  $X$ .

The characteristic  $E(X^k)$  is called the *k-th moment*; the characteristic  $\mu_k = E((X - EX)^k)$  is called the *k-th central moment* of a random variable  $X$ . What also comes in handy is the *k-th absolute moment*, given by  $E|X|^k$ .

From the definition it follows that for a continuous random variable  $X$ ,

$$E X^k = \int_{-\infty}^{\infty} x^k f_X(x) dx.$$



rearrange the expression:

$$\begin{aligned}
 & P \left[ \left| \frac{X}{n} - p \right| < 0.02 \right] \\
 &= P \left[ -0.02 < \frac{X}{n} - p < 0.02 \right] = \\
 &= P \left[ -0.02 \cdot n < X - pn < 0.02 \cdot n \right] = \\
 &= P \left[ \frac{-0.02 \cdot n}{\sqrt{np(1-p)}} < \frac{X - pn}{\sqrt{np(1-p)}} < \frac{0.02 \cdot n}{\sqrt{np(1-p)}} \right] = \\
 &= \Phi \left( \frac{0.02 \cdot n}{\sqrt{np(1-p)}} \right) - \Phi \left( -\frac{0.02 \cdot n}{\sqrt{np(1-p)}} \right) = \\
 &= 2\Phi \left( \frac{0.02 \cdot n}{\sqrt{np(1-p)}} \right) - 1,
 \end{aligned}$$

where  $\Phi$  is the distribution function of the normal distribution. Thus, let us solve the inequality

$$\begin{aligned}
 2\Phi \left( \frac{0.02 \cdot n}{\sqrt{np(1-p)}} \right) - 1 &\geq 0.99 \\
 \Phi \left( \frac{0.02 \cdot n}{\sqrt{np(1-p)}} \right) &\geq 0.995
 \end{aligned}$$

Since the distribution function is increasing, the last condition is equivalent to

$$\begin{aligned}
 \frac{0.02 \cdot n}{\sqrt{np(1-p)}} &\geq \Phi^{-1}(0.995) \\
 \frac{0.02 \cdot n}{\sqrt{np(1-p)}} &\geq 2.576 \\
 \sqrt{n} &\geq 50 \cdot 2.576 \cdot \underbrace{\sqrt{p(1-p)}}_{\leq \frac{1}{2}} \implies \\
 \implies n &\geq (25 \cdot 2.276)^2 \cdot 4147
 \end{aligned}$$

Here, we used the fact that the maximum of the function  $p(1-p)$  is  $\frac{1}{4}$ , and it is reached at  $p = \frac{1}{2}$ . We can see that if e. g.  $p \doteq 0.1$ , then  $\sqrt{p(1-p)} = 0.3$  and the value of the least  $n$  is lower. This accords with our expectations: for less popular parties, it suffices to have fewer respondents (if the agency estimates the gain of such party to be around 2 % without asking anybody, then the wanted precision is almost guaranteed). □

**10.I.11. Two-choice test.** Consider random vectors  $Y_1$  and  $Y_2$  all of whose components are pairwise independent random variables with normal distribution, and suppose that the components of vector  $Y_i$  have expected value  $\mu_i$ , and the variance  $\sigma$  is the same for all the components of both vectors.

Use the general linear model to test the hypothesis whether  $\mu_1 = \mu_2$ .

**Solution.** We will proceed quite similarly as in subsection 10.3.12 of the theoretical part. This time, we can write both

Similarly, for a discrete random variable  $X$  whose probability is concentrated into points  $x_i$ ,

$$E X^k = \sum_i x_i^k f_X(x_i).$$

The next theorem shows that all the moments completely describe the distribution of the random variable, as a rule.

For the sake of computations, it is advantageous to work with a power series in which the moments appear in the coefficients. Since the coefficients of the Taylor series of a function at a given point can be obtained using differentiation, it is easy to guess the right choice of such a function:

MOMENT GENERATING FUNCTION

Given a random variable  $X$ , consider the function  $M_X(t) : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$M_X(t) = E e^{tX} = \begin{cases} \sum_i e^{tx_i} f_X(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

If this expected value exists, the *moment generating function* of the random variable  $X$  can be discussed.

It is clear that this function  $M_X(t)$  is always analytic in the case of discrete random variables with finitely many values  $x_i$ .

**Theorem.** Let  $X$  be a random variable such that its analytic moment generating function on an interval  $(-a, a)$  exists. Then,  $M_X(t)$  is given on this interval by the absolutely convergent series

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E X^k.$$

If two random variables  $X$  and  $Y$  share their moment generating functions over a nontrivial interval  $(-a, a)$ , then their distribution functions coincide.

**PROOF.** The verification of the first statement is a simple exercise on the techniques of differential and integral calculus. In the case of discrete variables, there are either finite sums or absolutely and uniformly converging series. In the case of continuous variables, there are absolutely converging integrals. Thus, the limit process and the differentiation can be interchanged. Since  $\frac{d}{dt} e^{tx} = x e^{tx}$ , it is immediate that

$$\frac{d^k}{dt^k} M_X(t) = E X^k,$$

as expected.

The second claim is obvious for two discrete variables  $X$  and  $Y$  with only a finite number of values  $x_1, \dots, x_k$  for which either  $f_X(x_i) \neq 0$  or  $f_Y(x_i) \neq 0$ . Indeed, the functions  $e^{tx_i}$  are linearly independent functions and thus their coefficients in the common moment function

$$M(t) = e^{tx_1} f(x_1) + \dots + e^{tx_k} f(x_k)$$

must be the shared probability function values for both random variables  $X$  and  $Y$ .

vectors  $Y_i$  into one column, and we consider the model

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \sigma Z.$$

We will work with arithmetic means of the individual vectors  $\bar{Y}_1$  and  $\bar{Y}_2$ . Direct application of the general formula from theory gives the estimate  $b$  in the form

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} n_1 + n_2 & n_2 \\ n_2 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} n_1 \bar{Y}_1 + n_2 \bar{Y}_2 \\ n_2 \bar{Y}_2 \end{pmatrix} = \frac{1}{n_1 n_2} \begin{pmatrix} n_2 & -n_2 \\ -n_2 & n_1 + n_2 \end{pmatrix} \begin{pmatrix} n_1 \bar{Y}_1 + n_2 \bar{Y}_2 \\ n_2 \bar{Y}_2 \end{pmatrix} = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 - \bar{Y}_1 \end{pmatrix}$$

and for the matrix  $C = (X^T X)^{-1}$ , where  $X$  is the 2-column matrix with zeros and ones from our model, we have

$$C = \begin{pmatrix} \frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_2} \end{pmatrix}.$$

Thus, we test the hypothesis  $\mu_1 = \mu_2$ , which means that we test whether  $\beta_2 = 0$ . For this, it is suitable to use the statistic

$$T = \frac{\bar{Y}_2 - \bar{Y}_1}{S} \left( \frac{n_1 n_2}{n_1 + n_2} \right)^{\frac{1}{2}},$$

where the standard deviation  $S$  is substituted as

$$S^2 = \frac{1}{n_1 + n_2 - 2} \left( \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2 \right).$$

The distribution of this statistic is  $t_{n_1+n_2-2}$ , so the null hypothesis  $\mu_1 = \mu_2$  is rejected at level  $\alpha$  if we have

$$|T| \geq t_{n_1+n_2-2}(\alpha). \quad \square$$

**10.I.12.** In JZD<sup>1</sup> Tempo, the milk yield of their cows was measured during five days, the results being 15, 14, 13, 16 a 17 hectoliters. In JZD Boj, which had the same number of cows, they performed the same measurement during seven days, the results being 12, 16, 13, 15, 13, 11, 18 hectoliters.

- Find the 95% confidence interval for the milk yield of JZD Boj's cows, and the 95% confidence interval for the milk yield of JZD Tempo's cows.
- On the 5% level, test the hypothesis that both farms have cows of the same quality.

Suppose that the milk yield of the cows in each day is given by the normal distribution. Solve these problems assuming that there are no data from previous measurements, and then assuming that the previous measurements showed that the standard deviation was  $\sigma = 2$  hl.

<sup>1</sup>JZD — jednotné zemědělské družstvo — an agricultural cooperative farm, created by forced collectivization in 1950s in Czechoslovakia.

In the case of continuous variables  $X$  and  $Y$  sharing their generating function  $M(t)$ , the argument is more involved and an indication only is provided. Notice that  $M(t)$  is analytic and thus it is defined for all complex numbers  $t$ ,  $|t| < a$ .



In particular,

$$M(it) = \int_{-\infty}^{\infty} e^{itx} f(x) dx,$$

which is the inverse Fourier transform of  $f(x)$ , up to the constant multiple  $\sqrt{2\pi}$ , see 7.2.5 (on page 657). If this works for all  $t$ , then clearly  $f$  is obtained by the Fourier transform of  $(\sqrt{2\pi})^{-1} M(it)$  and thus must be the same for both  $X$  and  $Y$ . Further details, in particular covering general random variables, would need much more input from measure theory and Fourier analysis, and thus it is not provided here.  $\square$

It can be also shown that the assumptions of the theorem are true whenever both  $M_X(-a) < \infty$  and  $M_X(a) < \infty$ .

**10.2.37. Properties of moment function.** By the properties of the exponential functions, it is easy to compute the behaviour of the moment function under affine transformations and sums of independent random variables.



**Proposition.** Let  $a, b \in \mathbb{R}$  and  $X, Y$  be independent random variables with moment generating functions  $M_X(t)$  and  $M_Y(t)$ , respectively. Then, the moment generating functions of the random variables  $V = a + bX$  and  $W = X + Y$  are

$$\begin{aligned} M_{a+bX}(t) &= e^{at} M_X(bt) \\ M_{X+Y}(t) &= M_X(t) M_Y(t) \end{aligned}$$

**PROOF.** The first formula can be computed directly from the definition:

$$M_V(t) = \mathbb{E} e^{(a+bX)t} = \mathbb{E} e^{at} e^{(bt)X} = e^{at} M_X(bt).$$

As for the second formula, recall that  $e^{tX}$  and  $e^{tY}$  are independent variables. Use the fact that the expected value of the product of independent random variables equals the product of the expected values.

$$\begin{aligned} M_W(t) &= \mathbb{E} e^{t(X+Y)} = \mathbb{E} e^{tX} e^{tY} \\ &= \mathbb{E} e^{tX} \mathbb{E} e^{tY} = M_X(t) M_Y(t). \end{aligned} \quad \square$$

**10.2.38. Normal and binomial distributions.** As an illustrating example, compute the moment function of two random variables  $X \sim N(\mu, \sigma)$  and  $X \sim \text{Bi}(n, p)$ .

MOMENT GENERATING FUNCTION FOR  $N(\mu, \sigma)$

**Proposition.** If  $X \sim N(\mu, \sigma)$ , then

$$M_X(t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}.$$

In particular, it is an analytic function on all of  $\mathbb{R}$ .

**Solution.** First of all, let us compute the results for the known variance. In order to find the confidence interval, we use the statistic

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

which has standardized normal distribution (see 10.2.21).

Then, the confidence interval is (see 10.3.4)

$$\left( \bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2), \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2) \right),$$

where  $\alpha = 0,05$ . Now, it suffices to substitute the specific values. For JZD Tempo, we thus get the sample mean

$$\bar{X}_1 = \frac{15 + 14 + 13 + 16 + 17}{5} = 15,$$

and using appropriate software, we can learn that  $z(0.025) = 1.96$ , which gives the interval

$$\left( 15 - \frac{2}{\sqrt{5}}1.96, 15 + \frac{2}{\sqrt{5}}1.96 \right) \doteq (13.25; 16.75).$$

For JZD Boj, we get

$$\bar{X}_2 = \frac{12 + 16 + 13 + 15 + 13 + 11 + 18}{7} = 14,$$

so the 95% confidence interval for the milk yield of their cows is

$$(12.52; 15.48).$$

If the variance of the measurements is not known, we use the so-called sample variance for the estimate. In order to find the confidence interval, we use the statistic

$$T = \frac{\bar{X} - \mu}{S\sqrt{n}},$$

which has Student's distribution with  $n - 1$  degrees of freedom (see also 10.3.4). Then, we can analogously obtain the 95% confidence interval

$$\left( \bar{X} - \frac{S}{\sqrt{n}}t_{n-1}(\alpha/2), \bar{X} + \frac{S}{\sqrt{n}}t_{n-1}(\alpha/2) \right).$$

For the values of JZD Tempo, we get the sample variance

$$S_1^2 = \frac{0^2 + (-1)^2 + (-2)^2 + 1^2 + 2^2}{4} = 2.5,$$

i. e.,  $S \doteq 1.58$ . Further, we have  $t_4(0,025) \doteq 2,78$ , so the 95% confidence interval for JZD Tempo is

$$(13.03; 16.97).$$

For JZD Boj, we get the sample variance  $S_2^2 = 6$ , so the wanted confidence interval is

$$(11.73; 16.27).$$

b) If we compare the expected values of milk yield in both farms, then this is a comparison of the expected values of two independent choices from the normal distribution. In the case of unknown variances, we further assume that the variance is the same for both farms.

**PROOF.** Suppose  $Z \sim N(0, 1)$ . Then

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)} dx \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx \\ &= e^{\frac{t^2}{2}}, \end{aligned}$$

where use is made of the fact that in the last-but-one expression, for every fixed  $t$ , the density of a continuous random variable is integrated; hence this integral equals one.

Substitute the formula for the moment generating function  $M_{\mu+\sigma Z}$ , to obtain for  $X \sim N(\mu, \sigma)$  that

$$M_X(t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}},$$

again a function analytic over entire  $\mathbb{R}$ . □

In particular, the moments of  $Z$  of all orders exist. Substitute  $\frac{1}{2}t^2$  into the power series for the exponential function, and calculate them all:

$$\begin{aligned} M_Z(t) &= \sum_{k=0}^{\infty} \frac{1}{k!} \left( \frac{t^2}{2} \right)^k = \sum_{k=0}^{\infty} \frac{1}{k!2^k} t^{2k} = \\ &= 1 + 0t + \frac{1}{2}t^2 + 0t^3 + \frac{3}{4!}t^4 + \dots \end{aligned}$$

In particular, the expected value of  $Z$  is  $E Z = 0$ , and its variance is  $\text{var } Z = E Z^2 - (E Z)^2 = 1$ . Further, all moments of odd orders vanish,  $E Z^4 = 3$ , etc.

Hence the sum of independent normal distributions  $X \sim N(\mu, \sigma)$  and  $Y \sim N(\mu', \sigma')$  has again the normal distribution  $X + Y \sim N(\mu + \mu', \sigma + \sigma')$ .

Similarly, considering the discrete random variable  $X \sim \text{Bi}(n, p)$ ,

$$\begin{aligned} M_X(t) &= E e^{tX} = \sum_{k=0}^n (p e^t)^k \binom{n}{k} (1-p)^{n-k} \\ &= (p e^t + (1-p))^n = (p(e^t - 1) + 1)^n \\ &= 1 + npt + \frac{1}{2}(n(n-1)p^2 + np)t^2 + \dots \end{aligned}$$

is computed. Of course, the same can be computed even easier using the proposition 10.2.37 since  $X$  is the sum of  $n$  independent variables  $Y \sim A(p)$  with the Bernoulli distribution. Therefore,

$$E e^{tX} = (E e^{tY})^n = (p e^t + (1-p))^n.$$

Hence all the moments of the variable  $Y$  equal  $p$ . Therefore,  $E Y = p$ , while  $\text{var } Y = p(1-p)$ . From the moment function  $M_X(t)$ ,  $E X = np$  and  $\text{var } X = E X^2 - (E X)^2 = np(1-p)$ .

Thus, let us examine the hypothesis assuming the known variances  $\sigma_1^2 = \sigma_2^2 = 4$ . We use the statistic

$$\begin{aligned} U &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \\ &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1), \end{aligned}$$

where  $\mu_1$  and  $\mu_2$  are the unknown expected values of milk yield in the examined farms, and  $n_1, n_2$  are the numbers of measurements. This statistic has, as indicated, the standardized normal distribution. We reject the hypothesis at the 5% level if and only if the absolute value of the statistic  $U$  is greater than  $z_{0.025}$ , i. e., if and only if 0 does not lie in the 95% confidence interval for the difference of the expected values of milk yield in both farms. For the specific values, we get

$$U = \frac{15 - 14}{\sqrt{\frac{4}{5} + \frac{4}{7}}} \doteq 0.854.$$

Thus, we have  $|U| < z(0.025) = 1.96$ , so the hypothesis that the expected values of milk yield are the same in both farms is not rejected at the 5% level. The reached  $p$ -value of the test (see 10.3.9) is 39.4%, so we did not get much closer to rejecting the hypothesis (the probability that the value of the examined statistic is less than 0.854 provided the null hypothesis holds is 60.6%).

If we do not know the variances of the measurements but we know that they must be equal in both farms, we use the statistic

$$\begin{aligned} K &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \\ &= \frac{\bar{X}_1 - \bar{X}_2}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \end{aligned}$$

where

$$S_* = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

For the specific values, we get  $K \doteq 0.796$ ,  $|K| < t_{10}(0, 025) = 2.2281$ , so again, the null hypothesis is not rejected. The reached  $p$ -value of the test is 44.6%, which is even greater than in the above test.  $\square$

**10.I.13. Analyzing the variance of a simple sort.** For  $k \geq 2$  independent samples  $Y_i$  of size  $n_i$  from normal distributions with equal variance, use a linear model to test the hypothesis that all the expected values of individual samples are equal.

**Solution.** The technique is quite similar to that of the above exercise. The hypothesis to be tested is equivalent to stating that a submodel in which all the components of the random vector  $Y$  created by joining the given  $k$  vectors  $Y_i$  have the same expected value holds.

**10.2.39. Skewness and kurtosis.** Since the third central moment is given in terms of third powers of deviations from the expected value, it expresses to a certain extent the symmetry of the random variable distributed around the expected value. In descriptive statistics, we describe this by the coefficient of skewness. For random variables, we use similarly the characteristic

$$\gamma_1 = \frac{E(X - EX)^3}{(\sqrt{\text{var } X})^3},$$

which is called the *coefficient of skewness of a random variable*  $X$ .

Another commonly used characteristic is the *kurtosis* of a random variable  $X$ , defined as

$$\gamma_2 = \frac{E(X - EX)^4}{(\text{var } X)^2} - 3.$$

The standard normal distribution has third central moment equal to zero and the fourth one equal to 3. Thus, the kurtosis is standardized so that its value for the standard normal distribution is zero. For a general distribution, the kurtosis provides comparison to the normal distribution.

In practice, there are other standardizations of skewness coefficients and kurtosis.

**10.2.40. Law of large numbers.** Now, we can consider the key tools which connect probability and statistics. We start with the generalization of Bernoulli's theorem about the binomial distribution, discussed at the end of subsection 10.2.32.

The random variables  $\frac{1}{n}X_n$ , where  $X_n \sim \text{Bi}(n, p)$ , can be viewed as the arithmetic means of  $n$  independent variables with distribution  $A(p)$ , and Bernoulli's theorem then says that these means converge to  $p$  with probability 1.

Such a proposition holds in general. Independence of the variables is not needed, only the fact that  $\text{cov}(X_i, X_j) = 0$  guarantees that the variances sum up.

THE LAW OF LARGE NUMBERS

**Proposition.** Consider a sequence of pairwise uncorrelated random variables  $X_1, X_2, \dots$  which have the same finite expected value  $E X_i = \mu$ . Moreover, assume the variances are bounded, so that  $\text{var } X_i \leq C$ , for a fixed constant  $C$ . Then for any  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| < \varepsilon\right) = 1.$$

**PROOF.** By the use Chebyshev's inequality just as at the end of subsection 10.2.32,

$$\begin{aligned} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right) &\leq \frac{\text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu\right)}{\varepsilon^2} \\ &= \frac{\frac{1}{n^2} \sum_{i=1}^n \text{var } X_i}{\varepsilon^2} \leq \frac{C}{n\varepsilon^2}. \end{aligned}$$

Thus, the used model is of the form

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \sigma Z.$$

We can easily compute estimates for the expected values  $\mu_i$  using arithmetic means:

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Hence we get the estimate  $\hat{Y}_{ij} = \bar{Y}_i$ , so the residual sum of squares is of the form

$$RSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

The estimate of the common expected value in the considered submodel is

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^k n_i \bar{Y}_i,$$

where  $n = n_1 + \cdots + n_k$ , and the residual sum of squares in this submodel is

$$RSS^0 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2.$$

In the original model, there are  $k$  independent parameters  $\mu_i$ , while in the submodel, there is a single parameter  $\mu$ , so the tested statistic is of the form

$$F = \frac{(n - k) (RSS^0 - RSS)}{(k - 1) RSS}.$$

□

### J. Linear regression

We already met the linear regression in chapter three, subsection ???. Now, we will try to apply the same principle to problems which are often studied by statisticians.

One standard application of the linear regression is “laying a line” through given data. Thus, we have a sequence of measurements for which we record the values of two variables between which we anticipate linear dependency. A classical example is the dependency of a son’s height on his father’s height.

Thus, the probability  $P$  is bounded from below by

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| < \varepsilon\right) \geq 1 - \frac{C}{n\varepsilon^2},$$

which proves the proposition. □

Thus, existence and uniform boundedness of variances suffices for the means of pairwise uncorrelated variables  $X_i$  with zero expected value to converge (in probability) to zero.

**10.2.41. Central limit theorem.** The next goal is more ambitious. In addition to the law of large numbers, the stochastic properties of the fluctuation of the means  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  around the expected value  $\mu$  need to be understood. We focus first on the simplest case of sequences of independent and identically distributed random variables  $X_i$ . Then formulate a more general version of the theorem and provide only comments on the proofs.



Move to a sequence of normalized random variables  $X_i$ . Assume  $E X_i = 0$  and  $\text{var } X_i = 1$ . Assume further that the moment generating function  $M_X(t)$  exists and is shared by all the variables  $X_i$ .

The arithmetic means  $\frac{1}{n} \sum_{i=1}^n X_i$  are, of course, random variables with zero expected value, yet their variances are  $\frac{n}{n^2} = \frac{1}{n}$ . Thus, it is reasonable to renormalize them to

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i,$$

which are again standardized random variables. Their moment generating functions are (see proposition 10.2.37)

$$M_{S_n}(t) = E e^{\frac{t}{\sqrt{n}} \sum_i X_i} = \left( M_X\left(\frac{t}{\sqrt{n}}\right) \right)^n.$$

Since it is assumed that the variables  $X_i$  are standardized,

$$M_X\left(\frac{t}{\sqrt{n}}\right) = 1 + 0 \frac{t}{\sqrt{n}} + \frac{1}{2n} \frac{t^2}{n} + o\left(\frac{t^2}{n}\right),$$

where again  $o(G(n))$  is written for expressions which, when divided by  $G(n)$ , approach zero as  $n \rightarrow \infty$ , see subsection 6.1.12.

Thus, in the limit,

$$\lim_{n \rightarrow \infty} M_{S_n}(t) = \lim_{n \rightarrow \infty} \left( 1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right)^n = e^{\frac{t^2}{2}}.$$

This is just the moment generating function of the normal distribution  $Z \sim N(0, 1)$ , see the end of subsection 10.2.35. Thus, the standardized variables  $S_n$  asymptotically have the standard normal distribution.

We have thus proved a special version of the following fundamental theorem. Although the calculation is merely a manipulation of moment generating functions, many special cases were proved in different ways, providing explicit estimates for the speed of convergence, which of course is useful information in practice.

Notice that the following theorem does not require the probability distributions of the variables  $X_i$  to coincide!

**10.J.1.** Find the linear regression model for the dependence of  $Y$  on  $X$ , based on the following lists of measured data:  $X = [1, 4, 5, 7, 10]$ ,  $Y = [3, 7, 8, 12, 18]$ .

**Solution.** In order to find the parameters of the regression line, use the formulas derived in 10.3.12. Using the method of least squares, we try to minimize the distance of the vector  $b_1X + b_0$  from the vector  $Y$  with respect to the parameters  $b_1$  and  $b_0$ . This distance, as we know from chapter two, is minimal for the orthogonal projection of the vector  $Y$  onto the vector subspace generated by the vectors  $(1, \dots, 1)$  and  $(x_1, \dots, x_n)$ . For parameters  $b_0, b_1$  of the regression line  $Y = b_1X + b_0$ , we obtain

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{(1 - 5.4)(3 - 9.6) + \dots + (10 - 5.4)(18 - 9.6)}{((1 - 5.4)^2 + (4 - 5.4)^2 + (5 - 5.4)^2 + (7 - 5.4)^2 + (10 - 5.4)^2)} = 1.677.$$

Now, we can easily calculate the coefficient  $b_0$ :

$$b_0 = \bar{Y} - b_1\bar{x} = 0.5442.$$

Therefore, the wanted linear dependency is

$$Y = 1.677 \cdot X + 0.5442.$$

Note that in this model, the roles of the variables  $X$  and  $Y$  are totally equal. Using the same method, we could have obtained the dependency of  $X$  on  $Y$ :

$$X = 0.5867 \cdot Y - 0.2322. \quad \square$$

**Remark.** Think out why the linear regression model of the dependency of  $X$  on  $Y$  cannot be obtained by merely expressing  $X$  from the linear regression model of the dependency of  $Y$  on  $X$ .

**Remark.** In many real situations, the dependency of the variables is clearly given, if one of the variables is time, for example.

**10.J.2.** An orbital station has measured, at the same instant of five consecutive days, the following velocities of an unknown cosmic object (in km/s): 10, 11.4, 13.1, 15.8, and 18.7. Estimate the object's velocity on the tenth day.

**Solution.** Here, it is good to notice that the velocity does not change linearly with time (the acceleration is increasing). Thus, we can hypothesize that the object is being attracted to another one with the gravitational force. Then, its velocity would be a quadratic function of time. So let us use the method of least squares to lay a quadratic function (as precise as possible) through the measured data. The procedure is the same as if we made the linear regression of the vector  $v = (v_1, v_2, \dots, v_n)$  dependent on  $x = (x_1, \dots, x_n)$  and  $x^2 = (x_1^2, \dots, x_n^2)$ . This method is called *quadratic regression*. Thus, we are looking for a vector of parameters  $b = (b_0, b_1, b_2)$  so that the variable  $b_2x^2 + b_1x + b_0$  would

CENTRAL LIMIT THEOREM

**Theorem.** Consider a sequence of independent random variables  $X_i$  which have the same expected value  $E X_i = \mu$ , variance  $\text{var } X_i = \sigma^2 > 0$  and uniformly bounded third absolute moment  $E |X_i|^3 < C$ . Then, the distribution of the random variable

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)$$

satisfies

$$\lim_{n \rightarrow \infty} P(S_n < x) = \Phi(x),$$

where  $\Phi$  is the distribution function of the standard normal distribution.

Note that the central limit theorem gives a result on asymptotic behaviour which says that the distribution functions of certain variables approach the standard normal distribution. Such behaviour is called *convergence in distribution*. This type of convergence is weaker than convergence in probability.

The assumption that all  $X_i$  are independent and identically distributed was not fully exploited in the argumentation above. Only the knowledge of  $E X_i = 0$  and  $\text{var } X_i = 1$  was used. The assumption of the uniformly bounded third absolute moments of  $X_i$  can be used to prove the existence of the moment generating functions. The estimate  $E |X_i|^3$  can then be used to complete the above proof exactly as above.

There are many more general results. We mention at least the *Lyapunov's central limit theorem* formulated as follows:

Consider a sequence of random variables  $X_i$  with finite expected values  $\mu_i$  and variances  $\sigma_i^2$ . Write

$$s_n = \sqrt{\sum_{i=1}^n \sigma_i^2}$$

and assume for some  $\delta > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n E |X_i - \mu_i|^{2+\delta} = 0.$$

Then  $\frac{X_i - \mu_i}{s_n}$  converges in distribution to  $Z \sim N(0, 1)$ .

The previous version of the central limit theorem is derived by choosing  $\delta = 1$ . Then  $s_n = \sigma\sqrt{n}$  and the condition of the Lyapunov's theorem reads

$$0 \leq \lim_{n \rightarrow \infty} n^{-3/2} \sigma^{-3} \sum_{i=1}^n E |X_i|^3 \leq C\sigma^{-3} \lim_{n \rightarrow \infty} n^{-3/2+1} = 0.$$

**10.2.42. De Moivre-Laplace theorem.** Historically, the first formulated special case of the central limit theorem was that of variables  $Y_n$  with binomial distribution  $\text{Bi}(n, p)$ . They can be viewed as the sum of  $n$  independent variables  $X_i$  with Bernoulli distribution  $A(p)$ ,  $0 < p < 1$ . These variables have moment generating functions and  $E |X_i|^3 = p < 1$ .



estimate  $y$ . Let us build the matrix  $X$  of the values of independent variables:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 4 & 25 \end{pmatrix},$$

and the vector of parameters  $b = (b_0, b_1, b_2)$  can be computed by (1):

$$b = (X^T X)^{-1} X^T v \doteq (9.26; 0.47; 0.29).$$

Then, the wanted quadratic estimate is

$$v = 0.29x^2 + 0.47x + 9.26,$$

so the estimated velocity on the tenth day is approximately 42.96 km/s. In the model of classic linear regression, we would get

$$v = 2.18x + 7.26,$$

which yields 29.06 km/s for the tenth day. The difference between these estimates is quite large. This illustrates that analysis of the situation is a very important part of statistics.  $\square$

## K. Bayesian data analysis

**10.K.1.** Consider the Bernoulli process defined by a random variable  $X \sim \text{Bi}(n, \theta)$  with binomial distribution, and assume that the parameter  $\theta$  is a random variable with uniform distribution on the interval  $(0, 1)$ . We define the *success chance* in our process as the variable  $\gamma = \frac{\theta}{1-\theta}$ . What is the density of this variable  $\gamma$ ?

**Solution.** Intuitively, we can feel that the distribution is not uniform.

Denoting the wanted probability density by  $f(s)$ , we can use the relation between  $\theta$  and  $\gamma$  to compute  $\theta = \frac{\gamma}{1+\gamma}$ . In addition, we can immediately see that the probability density of  $\gamma$  is non-zero only for positive values of the variable. Now, we can formulate the statement as the requirement

$$(1) \quad \Theta = P(\theta < \Theta) = P(\gamma < \frac{\Gamma}{1+\Gamma}) = \int_0^\Gamma f(s) ds,$$

where  $\Gamma = \frac{\Theta}{1-\Theta}$ . However, the right-hand upper bound contains the changing limit  $\gamma$ , so we get the defining formula for  $f(s)$

$$f(s) = \left( \frac{s}{s+1} \right)' = \frac{1}{(s+1)^2}.$$

Indeed, the wanted density gives much higher probability to low values of the chance than to high ones.  $\square$

We could see in subsection 10.3.7 that when taking the Bayesian approach with binomial model of probability distribution of a random variable  $X \sim \text{Bi}(n, \theta)$ , then we are interested in its probability mass function  $f_X(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$ . This function can be viewed as the conditional probability  $P(\theta | X = k)$  for the uniform a priori probability distribution of the variable  $\theta$  on the interval  $(0, 1)$ . Thus, it is just

Thus, the central limit theorem says in this case that the random variables

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{X_i - p}{\sqrt{p(1-p)}} \right) = \frac{X - np}{\sqrt{np(1-p)}}$$

behave asymptotically as the standard normal distribution.

This can be formulated: the random variable  $X \sim \text{Bi}(n, p)$  behaves as the random variable with normal distribution  $N(np, np(1-p))$  as  $n$  increases.

This behaviour is demonstrated exactly in the illustration at the end of 10.2.21.

In practice, approximation of the binomial distribution by the normal distribution is usually considered appropriate if  $np(1-p) > 9$ .

We illustrate the result with a concrete example. Suppose it is desired to know what percentage of students like a given course, with an error of at most 5%. The number of people who like the course among  $n$  randomly chosen people should behave as the random variable  $X \sim \text{Bi}(n, p)$ . Further, suppose the result is desired to be correct with confidence (i.e., probability again) to at least 90%. Thus,

$$P\left( \left| \frac{1}{n} X - p \right| < 0.05 \right) \simeq 0.9$$

is desired by choosing a high enough number  $n$  of students to ask.

Approximate

$$\begin{aligned} 0.9 &\simeq P\left( \left| \frac{1}{n} X - p \right| < 0.05 \right) \\ &= P\left( -\frac{0.05n}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} < \frac{0.05n}{\sqrt{np(1-p)}} \right) \\ &\simeq \Phi\left( \frac{0.05n}{\sqrt{np(1-p)}} \right) - \Phi\left( -\frac{0.05n}{\sqrt{np(1-p)}} \right) \\ &= 2\Phi\left( \frac{0.05n}{\sqrt{np(1-p)}} \right) - 1, \end{aligned}$$

where the symmetry of the density function of the normal distribution is exploited. Thus,

$$\Phi\left( \frac{0.05n}{\sqrt{np(1-p)}} \right) \simeq \frac{1}{2}(1 + 0.9) = 0.95$$

is wanted. This leads to the choice (recall the definition of critical values  $z(\alpha)$  for a variable  $Z$  with standard normal distribution in subsection 10.2.30)

$$\Phi\left( \frac{0.05n}{\sqrt{np(1-p)}} \right) \simeq z(0.05) = 1.64485.$$

Since  $p(1-p)$  is at most  $\frac{1}{4}$ , the necessary number of students can be bounded by  $n > 270$ , independently of  $p$ .



the a posteriori probability distribution of  $\theta$  corresponding to the result  $X = k$  of the experiment. The following exercise concerns the general class of these probability distributions.

**10.K.2.** Find the basic characteristics of the so-called *beta-distribution*  $\text{fi}(a, b)$  with probability density of the form

$$f_Y = \begin{cases} C y^{a-1}(1-y)^{b-1} & y \in (0, 1) \\ 0 & \text{otherwise.} \end{cases}$$

**Solution.** The constant  $C$  must be chosen as the multiplicative inverse of the integral  $\int_0^1 y^{a-1}(1-y)^{b-1} dy$ , which is a function  $B(a, b)$ , known as *beta-function* in mathematical analysis and other sciences (e. g. physics). The function gamma, which generalizes the discrete values of factorial, emerges in the following calculation:

$$\begin{aligned} \Gamma(x)\Gamma(y) &= \int_0^\infty e^{-t} t^{x-1} dt \cdot \int_0^\infty e^{-s} s^{y-1} ds = \\ &= \int_0^\infty \int_0^\infty e^{-t-s} t^{x-1} s^{y-1} dt ds = \\ &\text{(substitution } t = rq, s = r(1-q)) \\ &= \int_{r=0}^\infty \int_{q=0}^1 e^{-r} (rq)^{x-1} (r(1-q))^{y-1} r dq dr = \\ &= \int_{r=0}^\infty e^{-r} r^{x+y-1} dr \cdot \int_{t=0}^1 q^{x-1} (1-q)^{y-1} dq = \\ &= \Gamma(x+y)B(x, y). \end{aligned}$$

Thus, we get the general formula

$$B(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

and it follows from properties of the gamma-function for positive integers  $a, b$  that

$$B(n-k+1, k+1) = \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1} \binom{n}{k}^{-1}.$$

We can directly compute that the expected value of the variable  $X \sim \text{fi}(a, b)$  with beta-distribution is (applying  $\Gamma(z+1) = z\Gamma(z)$ )

$$E X = \frac{B(a+1, b)}{B(a, b)} = \frac{a}{a+b}.$$

If  $a = b$ , then the expected value and median are  $\frac{1}{2}$ .

We can also directly calculate the variance

$$\text{var } X = E(X - E X)^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

Thus, for  $a = b$ , we get  $\text{var } X = \frac{1}{8a+4}$ , which shows that the variance decreases as  $a = b$  increases. For  $a = b = 1$ , we get the ordinary uniform distribution on the interval  $(0, 1)$ .  $\square$

**10.2.43. Important distributions.** In the sequel, we return to statistics. It should be of no surprise that we work with the characteristics of random vectors similar to the sample mean and variance, as well as relative quotients of such characteristics, etc. We consider several such cases.



Consider a random variable  $Z \sim N(0, 1)$ , and compute the density  $f_Y(x)$  of the random variable  $Y = Z^2$ . Clearly,  $f_Y(x) = 0$  for  $x \leq 0$ , while for positive  $x$ ,

$$\begin{aligned} F_Y(x) &= P(Y < x) = P(-\sqrt{x} < Z < \sqrt{x}) \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_0^x \frac{1}{\sqrt{2\pi}} t^{-1/2} e^{-t/2} dt. \end{aligned}$$

Differentiation leads to

$$f_Y(x) = \frac{d}{dx} F_Y(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}.$$

This distribution is called  $\chi^2$  with one degree of freedom, written  $Y \sim \chi^2$ .

We work with sums of such independent variables. All fall into a general class of distributions whose densities are of the form

$$f_X(x) = cx^{a-1} e^{-bx}$$

for  $x > 0$ , while  $f_X(x) = 0$  for non-positive  $x$ , i.e., the distribution  $\chi^2$  corresponds to the choice  $a = b = 1/2$ . This case is already thoroughly discussed as an example in subsection 10.2.20. Hence such a function is the density for the constant  $c = \frac{b^a}{\Gamma(a)}$ . Thus, it is the distribution  $\Gamma(a, b)$  with density, for positive  $x$ ,

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}.$$

In general, the  $k$ -th moment of such variable  $X$  is easily computed:

$$\begin{aligned} E X^k &= \int_0^\infty x^k \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} dx \\ &= \frac{\Gamma(a+k)}{\Gamma(a)b^k} \int_0^\infty \frac{b^{a+k}}{\Gamma(a+k)} x^{a-1+k} e^{-bx} dx \\ &= \frac{\Gamma(a+k)}{\Gamma(a)b^k}, \end{aligned}$$

since the integral of the density of  $\Gamma(a+k, b)$  in the last expression must be equal to one

In particular,  $E X = \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}$ , while

$$\text{var } X = \frac{\Gamma(a+2)}{b^2\Gamma(a)} - \frac{a^2}{b^2} = \frac{(a+1)a - a^2}{b^2} = \frac{a}{b^2}.$$

Similarly, the moment generating function can be computed for all values  $t$ ,  $-b < t < b$

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} dx = \\ &= \frac{b^a}{(b-t)^a} \int_0^\infty x^k \frac{(b-t)^a}{\Gamma(a)} x^{a-1} e^{-(b-t)x} dx = \\ &= \frac{b^a}{(b-t)^a}. \end{aligned}$$

**10.K.3.** In the situation as in the problem above the previous one, assume that the success chance  $\theta$  in the Bernoulli process is a random variable with probability distribution  $\text{fi}(a, b)$ . What is the probability distribution of the variable  $\gamma = \frac{\theta}{1-\theta}$ ? In what is it special when  $a = b = p$ ?

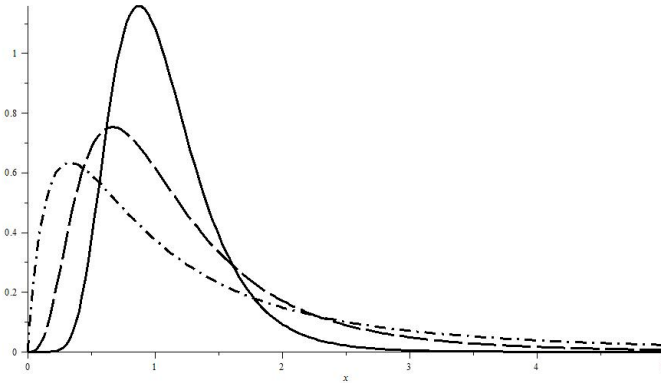
**Solution.** We have already discussed the special case with uniform distribution  $\text{fi}(1, 1)$ . Thus, we can continue with the equality  $\|1\|$ , where we used the form of this distribution. Now the left-hand side contains, instead of  $\Theta$ , the expression

$$\frac{1}{B(a, b)} \int_0^\Theta t^{a-1} (1-t)^{b-1} dt.$$

When differentiating, we must use the rule for differentiation of integral with variable upper bound. Thus, we get for the wanted density that

$$\begin{aligned} B(a, b)f(s) &= \left(\frac{s}{s+1}\right)^{a-1} \left(1 - \frac{s}{s+1}\right)^{b-1} \frac{1}{(s+1)^2} = \\ &= \left(\frac{s^{a-1}}{s+1}\right)^{a+b}. \end{aligned}$$

The picture shows the densities for  $a = b = p = 2, 5, 15$ .



This enforces the intuition that the same and not too small values of  $a = b = p$  correspond to the most probable value  $\theta = \frac{1}{2}$ , so the density of the chance is greatest around one. The higher  $p$ , the lower the variance of this variable.  $\square$

**10.K.4.** Show that the Bernoulli experiment, described by a random variable  $X \sim \text{Bi}(n, \theta)$ , and the a priori probability of a random variable  $\theta$  with beta-distribution, the a posteriori probability also has beta-distribution with suitable parameters which depend on the experiment results. What is the a posteriori expected value of  $\theta$  (i. e., the Bayesian point estimate of this random variable)?

**Solution.** As justified in subsection 10.3.7 of the theoretic part, the a posteriori probability density is, up to an appropriate constant, given as the product of the a priori probability density

$$g(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1}$$

and the probability of the examined variable  $X$  provided the value of  $\theta$  occurred. Thus, assuming  $k$  successes in the

Thus, for the sum of independent variables  $Y = X_1 + \dots + X_n$  with distributions  $X_i \sim \Gamma(a_i, b)$ , the moment generating function (for values  $|t| < b$ ) is obtained

$$M_Y(t) = \left(\frac{b}{b-t}\right)^{a_1+\dots+a_n},$$

that is,  $Y \sim \Gamma(a_1 + \dots + a_n, b)$ . It is essential that all of the gamma distributions share the same value of  $b$ .

As an immediate corollary, the density of the variable  $Y = Z_1^2 + \dots + Z_n^2$  is obtained, where  $Z_i \sim N(0, 1)$ . As just shown, this is the gamma distribution  $Y \sim \Gamma(n/2, 1/2)$ ; hence its density is

$$f_Y(x) = \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$

This special case of a gamma distribution is called  $\chi^2$  with  $n$  degrees of freedom. Usually, it is denoted by  $Y \sim \chi_n^2$ .

**10.2.44. The F-distribution.** In statistics, it is often wanted



to compare two sample variances, so we need to consider variables which are given as a quotient

$$U = \frac{X/k}{Y/m},$$

where  $X \sim \chi_k^2$  and  $Y \sim \chi_m^2$ .

Suppose  $f_X(x)$  and  $f_Y(y)$  are the densities of independent random variables  $X$  and  $Y$ . Suppose  $f_Y$  is non-zero only for positive values of  $x$ . Compute the distribution function of the random variable  $U = cX/Y$ , where  $c > 0$  is an arbitrary constant. By Fubini's theorem, the order of integration can be interchanged with respect to the individual variables.

$$\begin{aligned} F_U(u) &= P(X < (u/c)Y) = \int_0^\infty \int_{-\infty}^{uy/c} f_X(x) f_Y(y) dx dy \\ &= \int_0^\infty \left( \int_{-\infty}^u \frac{y}{c} f_X(ty/c) f_Y(y) dt \right) dy \\ &= \int_{-\infty}^u \left( \frac{1}{c} \int_0^\infty y f_X(ty/c) f_Y(y) dy \right) dt. \end{aligned}$$

This expression for  $F_U(u)$  shows that the density  $f_U$  of the random variable  $U$  equals

$$f_U(u) = \frac{1}{c} \int_0^\infty y f_X(uy/c) f_Y(y) dy.$$

Substitute the densities of the corresponding special gamma distributions for  $X \sim \chi_k^2$  and  $Y \sim \chi_m^2$ . Set  $c = m/k$ . The random variable  $U = \frac{X/k}{Y/m}$  has density  $f_U(u)$  equal to

$$\frac{(k/m)^{k/2}}{2^{(k+m)/2} \Gamma(k/2) \Gamma(m/2)} \int_0^\infty y^{(k+m)/2-1} e^{-y(1+ku/m)/2} dy.$$

The integrand in the latter integral is, up to the right constant multiple, the density of the distribution of a random variable  $Y \sim \Gamma((k+m)/2, (1+ku/m)/2)$ . Hence the multiple can be rescaled (notice  $u$  is constant there) in order to get

Bernoulli experiment, we get the a posteriori density (the sign used instead of equality denotes “proportional”)

$$\begin{aligned} g(\theta|X = k) &\propto P(X = k|\theta)g(\theta) \propto \\ &\propto \theta^k(1 - \theta)^{n-k}\theta^{a-1}(1 - \theta)^{b-1} = \\ &= \theta^{a+k-1}(1 - \theta)^{b+n-k-1}. \end{aligned}$$

Thus, we have indeed obtained the density (up to a constant, which we need not evaluate) of the a posteriori distribution for  $\theta$  with distribution  $B(a + k, b + n - k)$ .

Its a posteriori expected value is

$$\hat{\theta} = \frac{a + k}{a + b + n}.$$

For  $n$  and  $k$  approaching infinity so that  $k/n \rightarrow p$ , our a posteriori estimate also satisfies  $\hat{\theta} \rightarrow p$ . Thus, we can see that for large values of  $n$  and  $k$ , the observed fraction of successful experiments outweighs the a priori assumption. On the other hand, for small values, the a priori assumption is very important.  $\square$

**10.K.5.** We have data about accident rates for  $N = 20$  drivers in the last  $n = 10$  years (the  $k$ -th item corresponds to the number of years when the  $k$ -th driver had an accident):

$$0, 0, 2, 0, 0, 2, 2, 0, 6, 4, 3, 1, 1, 1, 0, 0, 5, 1, 1, 0.$$

We assume that the probabilities  $p_j$ ,  $j = 1, \dots, N$ , that the  $j$ -th driver has an accident in a given year are constants.

For each driver, estimate the probability that s/he has an accident in the following year (in order to determine the individual insurance fee, for instance). <sup>2</sup>

**Solution.** We introduce random variables  $X_{ij}$  with value 0 if the  $i$ -th driver has no accident in the  $j$ -th year, and 1 otherwise. The individual years are considered to be independent. Thus, we can assume that the random variables  $S_j = \sum_{i=1}^n X_{ji}$  that correspond to the number of accidents in all the  $n = 10$  years have distribution  $\text{Bi}(n, p_j)$ .

Of course, we could estimate the probabilities for all drivers altogether, i. e., using the arithmetic mean

$$\hat{p} = \frac{1}{N} \sum_{j=1}^n S_j \frac{1}{n} = \frac{1}{20} \frac{29}{10} = 0.145.$$

However, consider the homogeneity of the distribution of the variables  $X_j$ , they can hardly be accounted equal, so such estimate would be misleading.

On the other hand, the opposite extreme, i. e., a totally independent and individual estimate

$$\hat{p}_j = \frac{1}{n} S_j$$

is also inappropriate, since we surely do not want to set zero insurance fee until the first accident happens.

The realistic method is to use the same assumption for the a priori distribution of the probabilities  $p_j$  of accident

<sup>2</sup>This problem is taken from the contribution M. Friesl, Bayesovské odhady v některých modelech, published in: Analýza dat 2004/II (K. Kupka, ed.), Trilobyte Statistical Software, Pardubice, 2005, pp. 21-33.

the integral to evaluate to one. The density  $f_U(u)$  is then expressed as

$$\frac{\Gamma((k + m)/2)}{\Gamma(k/2)\Gamma(m/2)} \left(\frac{k}{m}\right)^{k/2} u^{k/2-1} \left(1 + \frac{k}{m}u\right)^{-(k+m)/2}.$$

This distribution is called the *Fisher-Snedecor distribution with  $k$  and  $m$  degrees of freedom*, or *F-distribution* in short.

**10.2.45. The t-distribution.** One encounters another useful distribution when examining the quotient of variables  $Z \sim N(0, 1)$  and  $\sqrt{X/n}$ . Here  $X \sim \chi_n^2$ . (We are interested in the quotient of  $Z$  and the standard deviation of some sample).

Compute first the distribution function of  $Y = \sqrt{X}$  (note that  $X$ , and hence  $Y$  as well, take only positive values with non-zero probability)

$$\begin{aligned} F_Y(y) &= P(\sqrt{X} < y) = P(X < y^2) \\ &= \int_0^{y^2} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} dx \\ &= \int_0^y \frac{1}{2^{n/2-1}\Gamma(n/2)} t^{n-1} e^{-t^2/2} dt. \end{aligned}$$

Hence the density of the random variable  $Y$  is

$$f_Y(y) = \frac{1}{2^{n/2-1}\Gamma(n/2)} y^{n-1} e^{-y^2/2}.$$

The same method can be used as in the previous subsection with the random variable  $U = cZ/Y$ , setting  $c = \sqrt{n}$ ,  $Y = \sqrt{X}$ . This leads to the random variable

$$T = \frac{Z}{\sqrt{X/n}}.$$

Similar computation as the one above yields that the density  $f_T$  satisfies

$$f_T(t) = \frac{\Gamma((n + 1)/2)}{\Gamma(n/2)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}.$$

This is called the *Student's t-distribution with  $n$  degrees of freedom*.

**10.2.46. Multidimensional normal distribution.** Consider



a random vector  $Z = (Z_1, \dots, Z_n)$  with independent components  $Z_i \sim N(0, 1)$ . Then its covariance matrix is equal to the unit matrix, i.e.,  $\text{var } Z = \mathbb{I}_n$ .

Random vectors are often encountered which are an affine transformation  $U = a + BZ$  of such a vector  $Z$ , where  $a$  is an arbitrary constant vector in  $\mathbb{R}^m$  and  $B$  is an  $m$ -by- $n$  constant matrix.

As derived in theorems 10.2.29 and 10.2.35, these random vectors have expected value  $E U = a$  and covariance matrix  $\text{var } U = V = B B^T$  (since the covariance matrix of  $Z$  is the identity matrix). Therefore, this covariance matrix is always positive-semidefinite.

The random vector  $U$  is said to have *multivariate normal distribution*  $N_m(a, V)$ .

rates of individual drivers. In practice, one often uses a model with the Poisson distribution  $\text{Po}(\lambda_j)$  for the  $j$ -th driver, with further assumptions about the distribution of the parameter  $\lambda$  among the drivers. We can also assume quite well (and simply) that the distribution is  $p_j \sim \text{fi}(a, b)$  with suitable parameters  $a, b$  which should reflect the cumulated results of all drivers. Thus, let us go this way.

We know from the above exercise that the a posteriori probability distribution will be  $(p_j | S_j = k) = \text{fi}(a + k, b + n - k)$ , so the corresponding expected value will be

$$\hat{p}_j^b = \frac{a + k}{a + b + n}.$$

Let us compare this estimate to the common estimate  $\hat{p}$  mentioned above and the individual estimate  $\hat{p}_j$ . We introduce the values  $p_0 = \frac{a}{a+b}$ , i. e., the expected value of the a priori common distribution for all drivers, and  $n_0 = a + b$ . We get

$$\hat{p}_j^b = \frac{(a+b)a}{(a+b+n)(a+b)} + \frac{nk}{(a+b+n)n} = \frac{n_0}{n_0+n} p_0 + \frac{n}{n_0+n} \hat{p}_j$$

which is a linear combination of the expected value  $p_0$  and the individual estimate  $\hat{p}_j$ .

Thus, it only remains to reasonably estimate the unknown parameters  $a, b$ . We know that

$$\begin{aligned} \mathbb{E} X_{ji} &= \mathbb{E} \mathbb{E}(X_{ji} | p) = \mathbb{E} p = p_0 \\ \frac{\mathbb{E} \text{var}(X_{ji} | p)}{\text{var} \mathbb{E}(X_{ji} | p)} &= \frac{\mathbb{E}(p(1-p))}{\text{var} p} = a + b = n_0 \end{aligned}$$

and the left-hand variables can be estimated directly.

$$\begin{aligned} \mathbb{E} X_{ij} &= \mathbb{E} \mathbb{E}(X_{ji} | p) \simeq \frac{1}{N} \sum_{j=1}^N \hat{p}_j \\ \mathbb{E} \text{var}(X_{ji} | p) &\simeq \frac{1}{N} \sum_{j=1}^N \left( \frac{n}{n-1} \hat{p}_j (1 - \hat{p}_j) \right) \\ \text{var} \mathbb{E}(X_{ji} | p) &\simeq s_{\hat{p}_j}^2 - \frac{1}{nN} \sum_{j=1}^N \left( \frac{n}{n-1} \hat{p}_j (1 - \hat{p}_j) \right), \end{aligned}$$

where  $s_{\hat{p}_j}^2$  denotes the sample variance between individual estimates (you can verify that the subtraction of the right-most expression guarantees that the last estimate is unbiased).

Since for the mentioned data, we get  $n_0 \simeq 3.8643$  and  $p_0 \simeq 0.1450$ , the Bayesian estimate of the individual probability of accidents is

$$\hat{p}_j^b = 0.154 \cdot 0.145 + 0.846 \cdot \hat{p}_j.$$

Thus, it is a combination of the confidence estimate  $\hat{p} = 0.145$  of the collective probability  $p_0$  with the individual (frequentist) estimate  $\hat{p}_j$ , which is measured from a small number  $n = 10$  of observations of one driver.  $\square$

For any multivariate normal distribution  $N_m(a, V)$ , consider again the affine transformation

$$W = c + DU$$

with a vector of constants  $c \in \mathbb{R}^k$  and an arbitrary  $k$ -by- $m$  constant matrix  $D$ . Direct calculation leads to

$$W = c + D(a + BZ) = (c + Da) + (DB)Z,$$

which is a random vector  $W \sim N_k(c + Da, DBB^T D^T)$ . Thus, the covariance matrix of the multivariate normal distribution behaves as a quadratic form with respect to affine transformations.

This straightforward idea shows that any linear combination of components of a random vector with the multivariate normal distribution is a random variable with the normal distribution. Similarly, any vector obtained by choosing only some of the components of the vector  $U$  is again a random vector with the multivariate normal distribution.

Note that when the random vector  $Z \sim N_n(0, \mathbb{I}_n)$  is transformed with an orthogonal matrix  $Q^T$ , then the joint distribution function of the random vector  $U = Q^T Z$  can be computed directly. If the transformation in coordinates as  $t = Q^T z$  is written, then its inverse is  $z = Qt$ , and the Jacobian of this transformation is equal to one. Hence (note that  $\sum_i z_i^2 = \sum_i t_i^2$ ). As in chapter 3, write  $z < u$  if all components satisfy  $z_i < u_i$ )

$$\begin{aligned} F_U(u) &= P(U_i < u_i, i = 1, \dots, n) = \\ &= \int \dots \int_{Q^T z < u} (2\pi)^{-n/2} e^{-\sum z_i^2/2} dz_1 \dots dz_n \\ &= \int \dots \int_{t < u} (2\pi)^{-n/2} e^{-\sum t_i^2/2} dt_1 \dots dt_n \\ &= \left( \int_{-\infty}^{u_1} (2\pi)^{-1/2} e^{-t_1^2/2} dt_1 \right) \dots \\ &\quad \dots \left( \int_{-\infty}^{u_n} (2\pi)^{-1/2} e^{-t_n^2/2} dt_n \right) \\ &= F_{U_1}(u_1) \dots F_{U_n}(u_n) \end{aligned}$$

Hence it directly follows that all components of the random vector  $U$  are again independent, and  $U \sim N_n(0, \mathbb{I}_n)$ .

### 3. Mathematical statistics



The data processing in mathematical statistics is based on quite sophisticated mathematics, but the actual methods are also very much dependent on the inputs from the diverse application fields.

Here we restrict attention to a few modest notes about statistical methods. We suggest curious readers to look for more detailed literature (which would reflect the field of application as well).

**L. Processing of multidimensional data**

Sometimes, we need to process multidimensional data: for each of  $n$  objects, we determine  $p$  characteristics. For instance, we can examine marks of several students in various subjects.

**10.L.1.** In his attempts, J.G.Mendel examined 10 plants of pea, and each was examined for the number of yellow and green seeds. The results of the experiment are summarized in the following table:

plant number	1	2	3	4	5	6	7	8	9	10
yellow seeds	25	32	14	70	24	20	32	44	50	44
green seeds	11	7	5	27	13	6	13	9	14	19
total seeds	36	39	19	97	37	26	45	53	64	63

It follows from the genetic models that the probability of occurrence of the yellow seed should be 0.75 (and 0.25 for green seed). At the asymptotic significance level 0.05, test the hypothesis that the results of Mendel’s experiments are in accordance with the model.

**Solution.** We test the hypothesis with the *Pearson’s chi-squared test*. We use the statistic

$$K = \sum_{j=1}^r \frac{(n_j - np_j)^2}{np_j},$$

where  $r$  is the number of sorting intervals (measurements; we have  $r = 10$ ),  $n_j$  is the actually measured frequency in the chosen sorting interval (we will count the number of yellow seeds),  $p_j$  is the expected frequency (by the assumed distribution), in our case,  $p_j = 0.75$ ,  $j = 1, \dots, 10$ . If the results of the experiment were really distributed as assumed in our model, we would have  $K \approx \chi^2(r - 1 - p)$ , where  $p$  is the number of estimated parameters in the assumed probability distribution. In our case, it is especially simple, since our model does not have any unknown parameters, so we have  $p = 0$  (the parameters may occur if, e. g., we assume that the probability distribution in our experiment is normal but with unknown variance and expected value; then we would have  $p = 2$ ). Thus,  $K \approx \chi^2(9)$ . The statistic is recommended to be used if the expected frequency of the characteristic in each of the sorting intervals is at least 5.

Let us write the data into a table:

$j$	$n_j$	$p_j$	$np_j$	$\frac{(n_j - np_j)^2}{np_j}$
1	25	0.75	27	0.148148
2	32	0.75	29, 25	0.258547
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	44	0.75	46.5	0.134409

The value of the statistic  $K$  for the given data is

$$K = 0.148148 + 0.258547 + \dots + 0.134409 = 1.797495.$$

**10.3.1. Introductory ideas.** In the descriptive statistics at the beginning of this chapter, we tried to equip the data sets with some characteristics which would carry essential information such as the sample mean, variance, etc.

Mathematical statistics works with some sample of a given data set, trying to describe to what extent the obtained statistics are relevant, or to find or improve an appropriate theoretical model for the behaviour of the entire data set, based on the collected data. The model is then used to either accept/reject a hypothesis about the data set, or to estimate the probability of an event that might happen in the future.

Consider an easy example: construct a wooden coin with heads and tails. Toss it  $n$  times, noting that it comes up heads  $k \leq n$  times. From this experiment, attempt to deduce the probability that the coin comes up heads in both of two more tosses.

There are two fundamental approaches to this problem. The first one is the classical statistics (or *frequentist statistics*). Build on the assumption that the individual tosses are independent and have the same probability of coming up heads, which is given by an objectively existing parameter  $\theta = p$  (unknown to us so far). Thus, the individual tosses are considered to be the realization of a random variable  $X$  with Bernoulli distribution. The probability of getting  $k$  heads out of  $n$  tosses is given by the binomial distribution, and it is expected that the “best possible” estimate for the parameter  $p$  is the ratio  $\theta = k/n$ . Usually the confidence of such an estimate is also wanted. This can be obtained from knowledge of the total number  $n$  of tosses and the asymptotic behaviour of the model as  $n$  increases. For instance, if the coin comes up heads 8 times out of 10, With a certain (mathematically estimated) confidence, it can be stated that the probability of the coin coming up heads in both of the subsequent tosses is  $0.8^2 = 0.64$ , a number much more than half.

The other possibility is quite different. Consider the parameter  $\theta$  to be a random variable from some chosen family of distributions, the collected data to be constants, and then try to deduce how to adjust the probability distribution of this random variable  $\theta$ . For example, suppose a (perhaps fair) coin is created, i.e. the expected value is (close to) 0.5, but the precision of the production ensures this only up to some small  $\varepsilon > 0$ . The experiment of tossing the coin  $n$  times allows the adjustment of the distribution within the preferred class. Thus, we build on some assumptions about the distribution and adjust the prior distribution in view of the experiment. This approach is called *Bayesian statistics*.

The first approach is based on the purely mathematical abstraction that probabilities are given by the frequencies of event occurrences in data samples which are so large that they can be approximated with infinite models. The central limit theorem can be used to estimate their confidence. From the statistical point of view, the probability is an idealization of the relative frequencies of the cases when an examined result

This value is less than  $\chi_{0.95}^2(9) = 16.9$ , so we do not reject the null hypothesis at level 0.05 (i. e., we do not refute the known genetic model).

□

occurs in many repeated experiments. This seeming advantage/rigor can become a disadvantage as soon as we are interested in the confidence of the data themselves and the suitability of the chosen experiment. The same problem occurs if we want to use frequentist statistics to estimate the probability of one or more outcomes of an experiment that is executed only once.

On the other hand, Bayesian statistics is an example of applying mathematics to “common sense” when we want to adjust our belief in light of new information.

It is interesting that, from the historical point of view, the first approach was the Bayesian one (for instance, Laplace and more as early as in the 18th century), which succumbed to frequentist statistics in the 20th century. In recent decades, Bayesian statistics has been returning, together with further new approaches.

**10.3.2. Random sample of a population.** Describe the first



approach of the above subsection. Thus, assume that there is a (huge) basic statistical set of  $N$  units, which is called the *population*, and each of the units has a numerical characteristic, i.e., there is a set of values  $(x_1, \dots, x_N)$ . From this set there is only a *sample* with values  $(X_1, \dots, X_n)$ .

In order to avoid the discussion of the actual size of the basic statistical set of  $N$  units, assume that the items of the sample are selected one by one and every item is always put back into the population. In addition, assume that every item has the same probability  $1/N$  of being chosen. This is a *random sample*.

The way of realizing the random sample can be viewed as working with a vector  $(X_1, \dots, X_n)$  of independent, identically distributed random variables. In particular, they have the same distribution function  $F_X(x)$  and moments

$$E X_i = \mu, \quad \text{var } X_i = \sigma^2.$$

The next step must be a derivation of the characteristics of the sample mean  $\bar{X}$  and the *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The following theorem explains why the coefficient  $\frac{1}{n-1}$  is selected instead of  $\frac{1}{n}$ , which is the case with  $s^2$  in subsection 10.1.6.

**Theorem.** *The sample mean  $\bar{X}$  computed from a random sample of size  $n$  whose distribution has finite expected value  $\mu$  and finite variance  $\sigma^2$  satisfies*

$$E \bar{X} = \mu, \quad \text{var } \bar{X} = \frac{1}{n} \sigma^2.$$

*The sample variance  $S^2$  satisfies*

$$E S^2 = \sigma^2.$$

**PROOF.** As derived in subsection 10.2.29,

$$E \bar{X} = \frac{1}{n} E \sum_{i=1}^n X_i = \frac{1}{n} n \mu = \mu.$$

Since the variables  $X_i$  are independent, additivity of variance can be used (derived in subsection 10.2.33). The variance behaves as a quadratic form with respect to multiplication by a scalar. Hence

$$\text{var } \bar{X} = \frac{1}{n^2} \text{var} \sum_{i=1}^n X_i = \frac{1}{n^2} n \sigma^2 = \frac{1}{n} \sigma^2.$$

The formula

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

can be verified simply by expanding the multiplications. Thus:

$$\begin{aligned} \text{E } s^2 &= \frac{1}{n} \text{E} \sum_{i=1}^n (X_i - \mu)^2 - \frac{n}{n} (\bar{X} - \mu)^2 = \\ &= \frac{1}{n} \sum_{i=1}^n \text{var } X_i - \text{var } \bar{X} = \\ &= \left(1 - \frac{1}{n}\right) \sigma^2. \end{aligned}$$

That is why the variance  $s^2$  is multiplied by the coefficient  $\frac{n}{n-1}$ , which leads just to the sample variance  $S^2$  and its expected value  $\sigma$ . Of course, this multiplication makes sense only if  $n \neq 1$ .  $\square$

### 10.3.3. Random sample of the normal distribution. In



practice, it is necessary to know not only the numerical characteristics of the sample mean and the variance, but also their total probability distributions. Of course, it can be derived only if the particular probability distribution of  $X_i$  is known. As a useful illustration, calculate the result for a random sample of the normal distribution.

It is already verified, as an example on properties of moment generating functions in 10.2.37, that the sum of random variables with the normal distribution results again in the normal distribution. Hence the sample mean must also have the normal distribution, and since both its expected value and variance are known,  $\bar{X} \sim N(\mu, \frac{1}{n} \sigma^2)$ .

The probability distribution of the sample variance is more complicated. Here, apply the ideas about multivariate normal distributions from subsection 10.2.37. Consider a vector  $Z$  of standardized normal variables

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

The same property holds for the vector  $U = Q^T Z$  with any orthogonal matrix  $Q$ . In addition,  $\sum_i U_i^2 = \sum_i X_i^2$ . Choose the matrix  $Q$  so that the first component  $U_1$  equals the sample mean  $\bar{Z}$ , up to a multiple. This means that the first column of  $Q$  is chosen as  $(\sqrt{n})^{-1}(1, \dots, 1)$ . Then  $U_1^2 = n\bar{Z}^2$ , so we

can compute:

$$\begin{aligned}\sum_{i=1}^n U_i^2 &= \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n (Z_i - \bar{Z})^2 + n\bar{Z}^2 \\ \sum_{i=2}^n U_i^2 &= \sum_{i=1}^n (Z_i - \bar{Z})^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2.\end{aligned}$$

Therefore, a multiple of the sample variance  $\frac{n-1}{\sigma^2}S^2$  is the sum of  $n-1$  squares of standardized normal variables, so the following theorem is proved:

**Theorem.** *Let  $(X_1, \dots, X_n)$  be a random sample from the  $N(\mu, \sigma^2)$  distribution. Then,  $\bar{X}$  and  $S^2$  are independent variables, and*

$$\bar{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right), \quad \frac{n-1}{\sigma^2}S^2 \sim \chi_{n-1}^2.$$

Hence, it immediately follows that the standardized sample mean

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S}$$

has Student's t-distribution with  $n-1$  degrees of freedom.

**10.3.4. Point and interval estimates.** Now, we have every-



thing needed to estimate the parameter values in the context of frequentist statistics. Here is a simple example. Suppose there are 500 students enrolled in a course, each of which has a certain degree of satisfaction with the course, expressed as an integer in the range 1 through 10. It may be assumed that the satisfactions  $X_i$  of the students are approximated by a random variable with distribution  $N(\mu, \sigma^2)$ . Further, suppose a detailed earlier survey showed that  $\mu = 6$ ,  $\sigma = 2$ .

In the current semester, 15 students are asked about their opinion about the course, as rumor has it that the evaluation of the new lecturer might be quite different. The results show that 2 students vote 3, 3 vote 4, 3 vote 5, 5 vote 6, and 2 vote 7. Altogether, the sample mean is  $\bar{X} = 5.133$  and the sample variance is  $S^2 = 1.695$ .

By assumptions,  $\bar{X} \sim N(\mu, \sigma^2/n)$ , so  $Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$ . In order to express the confidence of the estimate, compute the interval which contains the estimated parameter with an a priori fixed probability  $100(1-\alpha)\%$ . We talk about a confidence level  $\alpha$ ,  $0 < \alpha < 1$ . Consider  $\mu$  to be the unknown parameter, while the variance can be assumed (be it correct or not) to remain unchanged. It follows that

$$\begin{aligned}1 - \alpha &= P(|Z| < z(\alpha/2)) = P\left(\left|\sqrt{n} \frac{\bar{X} - \mu}{\sigma}\right| < z(\alpha/2)\right) \\ &= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2)\right),\end{aligned}$$

where  $z(\alpha/2)$  means the critical value, cf. 10.2.30. Thus, an interval is found whose endpoints are random variables and which contains the estimated parameter  $\mu$  with an a priori fixed probability. The middle point of this interval is called the *point estimate* for parameter  $\mu$ ; the whole interval is called the *interval estimate*. We can also say that at the confidence



level  $\alpha$ , the estimated parameter  $\mu$  is or is not different from another value  $\mu_0$ . Suppose for instance, the data and levels are  $\alpha = 0.05$  and  $\alpha = 0.1$ . Respectively we obtain the intervals

$$\mu \in (4.121, 6.145), \quad \mu \in (4.284, 5.983).$$

Considering the confidence level of 5%, we cannot affirm that the opinion of students are worse compared to the previous year because the mentioned interval also contains the value  $\mu_0 = 6$ . We can conclude this if we take the confidence level of 10% since the value  $\mu_0 = 6$  no more lies in the corresponding interval.

On the other hand, if it is assumed that the other (worse) lecturer causes the variance of the answers to change as well (for instance, the students might agree more on the bad assessment), we proceed differently. Instead of the standardized variable  $Z$ , deal in a similar way with the variable

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S}.$$

As seen, this random variable has probability distribution  $T \sim t_{n-1}$ , where  $n = 15$  in this case. This leads to the interval estimate

$$\bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2).$$

Substitute the data at levels  $\alpha = 0.05$  and  $\alpha = 0.03$  respectively, to obtain

$$\mu \in (4.412, 5.854), \quad \mu \in (4.321, 5.945).$$

Therefore, at the confidence level of 3%, the opinion seems to have become worse. This corresponds to our intuition that the sample deviation  $S = 1.302$ , which is significantly smaller than  $\sigma = 2$  from the previous case, should be essential for our thinking.

**10.3.5. Likelihood of estimates.** From the mathematical



point of view, interval and point estimates are simple and easy to understand. It is much worse with their practical interpretation because it is problematic to verify all assumptions about randomness of the sample. With more complicated cases, we consider problems with the “likelihood” of our estimates.

As mathematicians, we can avoid the practical problem by defining the missing concept. In general, one works with a random sample of size  $n$ . Implicitly it is assumed that there are independent random variables  $X_i$  with the same probability distribution which depends on an unknown parameter  $\theta$  (a vector in general).

We are trying to find a *sample statistic*  $T$ , i.e., a function of the random variables  $X_1, X_2, \dots$  which, in a mathematical sense, estimates the actual value of the parameter  $\theta$ .  $T$  is said to be an *unbiased estimator* of  $\theta$  if and only if  $E T = \theta$ . The expected value  $E(T - \theta)$  is called the *bias of estimator*  $T$ .

The asymptotic behaviour of the estimator, that is, what it does as  $n$  goes to infinity is often of interest.  $T = T(n)$  is

said to be a *consistent estimator* of the parameter  $\theta$  if and only if  $T(n)$  converges in probability to  $\theta$ , i.e., for every  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|T(n) - \theta| < \varepsilon) = 1.$$

The Chebyshev's inequality immediately yields

$$P(|T(n) - \mathbb{E}T(n)| < \varepsilon) \geq 1 - \frac{\text{var } T(n)}{\varepsilon^2}.$$

Assuming  $\lim_{n \rightarrow \infty} \mathbb{E}T(n) = \theta$ , then, for sufficiently large values of  $n$ ,

$$P(|T(n) - \theta| < 2\varepsilon) \geq P(|T(n) - \mathbb{E}T(n)| < \varepsilon) \geq 1 - \frac{\text{var } T(n)}{\varepsilon^2}.$$

A useful proposition is thus proved :

**Theorem.** Assume that  $\lim_{n \rightarrow \infty} \mathbb{E}T(n) = \theta$  and  $\lim_{n \rightarrow \infty} \text{var } T(n) = 0$ . Then,  $T(n)$  is a consistent estimator of  $\theta$ .

As a simple example, we can illustrate this theorem on variance:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2.$$

Since it is known from subsection 10.3.2 that  $S^2$  is an unbiased estimator, it follows that  $\hat{\sigma}^2$  is not. However,  $\lim_{n \rightarrow \infty} \hat{\sigma}^2 = \sigma^2$ , and it can be calculated that

$$\lim_{n \rightarrow \infty} \text{var } \hat{\sigma}^2 = \lim_{n \rightarrow \infty} \text{var } S^2 = \lim_{n \rightarrow \infty} \frac{2\sigma}{n-1} = 0.$$

Therefore, the statistic  $s^2$  is a consistent estimator of the variance.

It is apparent that there may be more unbiased estimators for a given parameter. For instance, it is already shown that the arithmetic mean  $\bar{X}$  is an unbiased estimator of the expected value  $\theta$  of random variables  $X_i$ . The value  $X_1$  is, of course, an unbiased estimator of  $\theta$  as well. We wish to find the best estimator  $T$  in the class of considered statistics, which are unbiased or consistent. Consider as best the one whose variance is as small as possible. Recall that the variance of a vector statistic  $T$  is given by the corresponding covariance matrix, which is, in case of independent components, a diagonal matrix with the individual variances of the components on the diagonal. We have already defined inequalities between positive-definite matrices.

**10.3.6. Maximum likelihood.** Assume that the density function of the components of the sample is given by a function  $f(x, \theta)$  which depends on an unknown parameter  $\theta$  (a vector in general). By the assumed independence, the joint density of the vector  $(X_1, \dots, X_n)$  is equal to the product:

$$f(x_1, \dots, x_n, \theta) = f(x_1, \theta) \cdots f(x_n, \theta),$$

which is called the *likelihood function*.

We are interested in the value  $\hat{\theta}$  which maximizes the likelihood function on the set of all admissible values of the parameter. In the discrete case, this means choosing the parameter for which the obtained sample has the greatest probability.



Usually, it is more efficient to work with the *log-likelihood function*

$$\ell(x_1, \dots, x_n, \theta) = \ln f(x_1, \dots, x_n, \theta) = \sum_{i=1}^n \ln f(x_i, \theta).$$

Since the  $\ln$  function is strictly increasing, maximization of the log-likelihood function is equivalent to maximization of the original likelihood function. If, for some input, it happens that  $f(x_1, \dots, x_n, \theta) = 0$ , set  $\ell(x_1, \dots, x_n, \theta) = -\infty$ .

In the case of discrete random variables, use the same definition with probability mass function instead of the density, i.e.,

$$\ell(x_1, \dots, x_n, \theta) = \sum_{i=1}^n \ln(P(X_i = x_i|\theta)).$$

We can illustrate the principle on a random sample from the normal distribution  $N(\mu, \sigma^2)$  with size  $n$ . The unknown parameters are  $\mu$  or  $\sigma$ , or both. The considered density is

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Take logarithms of both sides, to obtain

$$\ell(x, \mu, \sigma) = -n\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

The maximum can be found using differentiation (note that  $\sigma^2$  is treated as a symbol for a variable):

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (-2)(x_i - \mu) = \frac{1}{\sigma^2} \left( -n\mu + \sum_{i=1}^n x_i \right) \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 = \\ &= \frac{1}{2(\sigma^2)^2} \left( -n\sigma^2 + \sum_{i=1}^n (x_i - \mu)^2 \right). \end{aligned}$$

Thus, the only critical points are given by  $\hat{\mu} = \bar{X}$  and  $\hat{\sigma}^2 = s^2$ . Substitute these values into the matrix of second derivatives, to obtain the Hessian of  $\ell$ :

$$\begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}.$$

Finally, this is the required maximum, and since there is only one critical point, it must be the global maximum (think about the details of this argument!).

Thus it is verified that the expected value and the variance are the most likely estimates for  $\mu$  and  $\sigma$ , as already used.

### 10.3.7. Bayesian estimates.

We return to the example from subsection 10.3.4, now from the point of view of Bayesian statistics. This totally reverses the approach: the collected data  $X_1, \dots, X_{15}$  (i.e., the points which express how much each student is satisfied, using the scale 1 through 10) are treated as constants. On the other hand, the estimated parameter  $\mu$  (the expected value of the points of satisfaction), is viewed as the random variable



whose distribution we wish to estimate. Let us look at the general principle first (and come back to this example soon).

For this purpose, let us interpret Bayes' formula for conditional probability on the level of probability mass functions or probability densities, in the following way: If a vector  $(X, \Theta)$  has joint density  $f(x, \theta)$ , then the conditional probability of a component  $\Theta$ , given by  $X = x$ , is defined as the density

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)},$$

where  $f(x)$  and  $g(\theta)$  are the marginal probability densities. (In the above example,  $x$  is 15-dimensional vector coming from the multidimensional normal distribution while  $\theta = \mu$  is scalar.)

Thus, given the *a priori* probability density  $g(\theta)$  of the estimated parameter  $\theta$  and the probability density  $f(x|\theta)$  (in the above example,  $\theta = \mu$  is the expected value parameter of the distribution), the formula to compute the *a posteriori* probability density  $g(\theta|x)$  can be used, based on the collected data. Indeed, we do not need to know  $f(x)$  for the following reason: we have to view  $f(x)$  as a constant independent of  $\theta$  and thus the proper density is obtained from  $f(x|\theta)g(\theta)$  by multiplying with a uniquely given constant in the end. Thus, during the computation, it is sufficient to be precise "up to a constant multiple". For this purpose, use the notation  $Q \propto R$ , meaning that there is a constant  $C$  such that the expressions  $Q$  and  $R$  satisfy  $Q = CR$ .

We shall illustrate this procedure on a more explicit example. In order to be as close as possible to the ideas from subsection 10.3.4, work with normal distributions  $N(\mu, \sigma^2)$ . Suppose that the satisfaction of individual students in particular lectures is a random variable  $X \sim N(\theta, \sigma^2)$ , while the parameter  $\theta$  reached by the particular lecturers is a random variable  $\theta \sim N(a, b)$ .

Compute, (up to a constant multiple, ignoring all multiplicative components which do not include any  $\theta$ ),

$$\begin{aligned} g(\theta|x) &\propto f(x|\theta)g(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-a)^2}{2b^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\theta^2\left(\frac{1}{\sigma^2} + \frac{1}{b^2}\right) - 2\theta\left(\frac{x}{\sigma^2} + \frac{a}{b^2}\right)\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\theta - \frac{b^2x + \sigma^2a}{\sigma^2b^2 + b^2 + \sigma^2}\right)^2 \left(\frac{b^2\sigma^2}{b^2 + \sigma^2}\right)^{-1}\right). \end{aligned}$$

This proves already that the distribution for  $\theta$  is

$$\theta \sim N\left(\frac{b^2}{b^2 + \sigma^2}x + \frac{\sigma^2}{b^2 + \sigma^2}a, \frac{b^2\sigma^2}{b^2 + \sigma^2}\right).$$

This result can be interpreted so that if the parameters  $a, b, \sigma$  are known from long-run evaluation of surveys and the opinion of another student is learned, then the *a priori* opinion about the parameters for an individual lecture can be adjusted.

In the resulting estimate, the expected value is given by the weighted average of the found value  $x$  and the *a priori* assumed expected value  $a$ , in dependence on the standard deviations  $\sigma$  and  $b$ .

**10.3.8. Interpretation in Bayesian statistics.** We follow the ideas from the previous subsection, compared to the frequentist interpretation from 10.3.4. It may seem odd that a single query can influence an opinion so much.



For  $\sigma \rightarrow 0$ , the relevance of a single opinion is still increasing, and this corresponds to a 100% relevance of  $x$  in the case  $\sigma = 0$ . This is in accordance with the interpretation that Bayesian statistics is the probability extension of the standard discrete mathematical logic. If the variance  $\sigma$  is close to zero, then it is almost certain that the opinion of any student precisely describes the opinion of the whole population.

In subsection 10.3.4, we worked with the sample mean  $\bar{X}$  of the collected data. This can be used in the previous calculation, since the mean also has a normal distribution, too. The expected value is the same, and the only difference is that  $\sigma^2/n$  is substituted instead of  $\sigma^2$ . To facilitate the notation, define the constant

$$c_n = \frac{nb^2}{nb^2 + \sigma^2}.$$

The a posteriori estimate for  $\theta$  based on the found sample mean  $\bar{X}$  has the distribution with parameters

$$\theta \sim N(c_n \bar{X} + (1 - c_n)a, c_n \sigma^2/n).$$

As could be expected, for increasing  $n$ , the expected value of the distribution for  $\theta$  approaches the sample mean, and its variance approaches zero. In other words, the higher the value of  $n$ , the closer is the point estimate from the frequentist point of view.

A contribution of the Bayesian approach is that if the estimated distribution is used, questions of the kind: “What is the probability that the new lecturer is worse than the old one?” can be answered. Use the same data as in 10.3.4 and supplement the necessary a priori data. Assume that the lecturers are assessed quite well (otherwise, they would probably not be teaching at the university at all). For concreteness, select the a priori distribution with parameters  $a = 7.5$ ,  $b = 2.5$ , and the standard deviation with  $\sigma = 2$ . Continue with  $n = 15$  and the sample mean of 5.133. Substitute this data, to get the a posteriori estimate for the distribution

$$\theta \sim N(5.230, 0.256).$$

We are interested in  $P(\theta < 6)$ . This is computed by evaluating the distribution function of the corresponding normal distribution for the input value 6 (Excel is capable of this, too). The answer is approximately 93.6 %. This is similar to the material in subsection 10.3.4, where the known variance is assumed constant.

Note the influence of the a priori assumption about the distribution of the parameter  $\theta$  for all lecturers. To a certain extent, this reflects a faith that the lecturers are rather good. If a statistician has a reason for assuming that the actual expected value  $a$  for a specific lecturer is shifted, say  $a = 6$  as in the survey about the previous lecturer, (this can be caused,

for example, by the fact that the lecture is hard and unpopular), then the probability of his actual parameter being less than 6 would be approximately 95.0 %. (If the expected value is considered to be significantly worse only when below 5.5, then the value would be only approximately 75 %). When substituting  $a = 5$ , the value is already 96.8 %. The variance  $b^2$  is also important. For instance, the a priori estimate  $a = 6$ ,  $b = 3.5$  leads to probability 95.2 %.

In the above discussion, another very important point is touched on – *sensitivity analysis*. It would seem desirable that a small change of the a priori assumption has only a small effect on the a posteriori result. It appears that this is so in this example; however, we omit further discussion here.

The same model with exponential distributions is used in practice when judging the relevance of the output of an IQ test of an individual person. It can also be used for another similar exam where it is expected that the normal distribution approximates well the probability distribution of the results. In both cases, there is an a priori assumption to which group he/she should belong. Other good examples (with different distributions) are practical problems from insurance industry, where it is purposeful to estimate the parameters so that both the effects of the experiment upon an individual item and the altogether expectations over the population are included.

**10.3.9. Notes on hypothesis testing.** We return to deciding whether a given event does or does not occur in the context of frequentist statistics. We build on the approach from interval estimates, as presented above.



Thus, consider a random vector  $X = (X_1, \dots, X_n)$  (the result of a random sample), whose joint distribution function is  $F_X(x)$ . A *hypothesis* is an arbitrary statement about the distribution which is determined by this distribution function. Usually, one formulates two hypothesis, denoted  $H_0$  and  $H_A$ . The former is traditionally called *null hypothesis*, and the latter is called *alternative hypothesis*. The result of the test is then a decision based on a concrete realization of the random vector  $X$  (a *test*) whether the hypothesis  $H_0$  is to be rejected or not in favor of the hypothesis  $H_A$ .

During this process, two types of errors may occur. *Type I error* occurs when  $H_0$  is rejected even though it is true. *Type II error* occurs when  $H_0$  is not rejected although it is false. The decision procedure of a frequentist statistician consists of selecting the *critical region*  $W$ , i.e., the set of test results when the hypothesis is rejected. The size of the critical region is chosen so that a true hypothesis is rejected with probability not greater than  $\alpha$ . This means that a fixed bound for the probability of the type I error is required: the *significance level*  $\alpha$ . The most common choices are  $\alpha = 0.05$  or  $\alpha = 0.01$ . It is also useful in practice to determine the least possible significance level  $p$  for which the hypothesis is rejected; the *p-value* of the test.

It remains to find a reasonable procedure for choosing the critical range. This should hopefully be done so that the

type II error occurs as rarely as possible. Usually, it is convenient to consider the likelihood function  $f(x, \theta)$ , defined for a random vector  $X$  in subsection 10.3.6. For the sake of simplicity, assume there is a one-dimensional parameter  $\theta$ , and formulate the null hypothesis as  $X$  being given by the function  $f(x, \theta_0)$ , while the alternative hypothesis is given by the distribution  $f(x, \theta_1)$  for fixed distinct values  $\theta_0$  and  $\theta_1$ . Ideas about rejecting or accepting the hypotheses suggest that when substituting the values of a specific test into the likelihood function, the hypothesis can be accepted if  $f(x, \theta_0)$  is much greater than  $f(x, \theta_1)$ .

This suggests considering, for each constant  $c > 0$ , the critical range

$$W_c = \{x; f(x, \theta_1) \geq cf(x, \theta_0)\}.$$

Having chosen the significance level, choose  $c$  so that

$$\int_{W_c} f(x, \theta_0) = \alpha.$$

This guarantees that for the test result  $x \in W_c$ , when  $H_0$  is valid, the type I error occurs with at most the prescribed probability. This can also be guaranteed by other critical ranges  $W$  which also satisfy

$$\int_W f(x, \theta_0) = \alpha.$$

On the other hand, type II errors are also of interest. That is, it is desired to maximize the probability of  $H_A$  over the critical range. Thus, consider the difference

$$D = \int_{W_c} f(x, \theta_1) - \int_W f(x, \theta_1)$$

for arbitrary  $W$  as above. The regions over which integration is carried out, can be divided into the common part  $W \cap W_c$  and the remaining set differences. The contributions of the common part are subtracted, and there remains

$$D = \int_{W_c \setminus W} f(x, \theta_1) - \int_{W \setminus W_c} f(x, \theta_1).$$

Using the definition of the critical range  $W_c$ , (again, put back the same integrals over the common part)

$$\begin{aligned} D &\geq c \int_{W_c \setminus W} f(x, \theta_0) - c \int_{W \setminus W_c} f(x, \theta_0) = \\ &= c \int_{W_c} f(x, \theta_0) - c \int_W f(x, \theta_0) = c\alpha - c\alpha = 0. \end{aligned}$$

Thus is proved an important statement, the *Neyman–Pearson lemma*:

**Proposition.** *Under the above assumptions,  $W_c$  is the optimal critical range which minimizes occurrence of the type II error at a given significance level.*

**10.3.10. Example.** The interval estimate, as illustrated on an example in subsection 10.3.4, is a special case of hypothesis testing, when  $H_0$  had the form “the expected value of the satisfaction with the course remained  $\mu_0$ ”, while  $H_A$  says that it is equal to a different value  $\mu_1$ . The general procedure mentioned above leads in this case to the critical range given by



$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \right| \geq z(\alpha/2).$$

Note that in the definition of the critical range, the actual value  $\mu_1$  is not essential. In the context of classical probability, the decision at a given level  $\alpha$  whether or not there is a change to the expected value  $\mu$  is thus formalized.

To test only whether the satisfaction is decreased, assume beforehand that  $\mu_1 < \mu_0$ . We analyze this case thoroughly: The critical range from the Neyman–Pearson lemma is determined by the inequality

$$\frac{f(x, \mu_1, \sigma^2)}{f(x, \mu_0, \sigma^2)} = e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n ((x_i - \mu_1)^2 - (x_i - \mu_0)^2)} \geq c.$$

Take logarithms and rearrange to obtain

$$2\bar{x}(\mu_1 - \mu_0) - (\mu_1^2 - \mu_0^2) \geq \frac{2\sigma^2}{n} \ln c.$$

Since  $\mu_1 < \mu_0$ , it follows that

$$\bar{x} \leq \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{n(\mu_1 - \mu_0)} \ln c = y.$$

For a given level  $\alpha$ , the constant  $c$ , and thereby the decisive parameter  $y$ , are determined so that, under the assumption that  $H_0$  is true,

$$\alpha = P(\bar{X} \leq y) = P\left(\frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \leq \frac{y - \mu_0}{\sigma} \sqrt{n}\right).$$

By assuming that  $H_0$  is true,

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \sim N(0, 1),$$

so the requirement means choosing  $Z \leq -z(\alpha)$ , which determines uniquely the optimal  $W_c$ .

Note that this critical region is independent of the chosen value  $\mu_1$ , and the actual value for  $y$  did not have to be expressed at all. It was only essential to assume that  $\mu_1 < \mu_0$ .

In the illustrative example from subsection 10.3.4,  $H_0 : \mu = 6$ , and the alternative hypothesis is  $H_A : \mu < 6$ . The variance is  $\sigma^2 = 4$ . The test with  $n = 15$  yielded  $\bar{x} = 5.133$ . Substitute this, to get the value  $z = \frac{5.133 - 6}{2} \sqrt{15} = -1.678$ , while  $-z(0.05) = -1.645$ .



Therefore, if we are testing whether the new teacher is even worth than the previous one, we reject the hypothesis at the level of 5 %, deducing that the students’ opinions are really worse.

If, for the critical range, the union of the critical ranges for the cases  $\mu_1 < \mu_0$  and  $\mu_1 > \mu_0$  are chosen, the same results as for the interval estimate are obtained, as mentioned above.



We remark that in the Bayesian approach, it is also possible to accept or reject hypotheses in a direct connection to the a posteriori probability of events, as was, to certain extent, indicated in subsection 10.3.8 where our specific example is interpreted.

**10.3.11. Linear models.** As is usual in the analysis of mathematical problems, either we deal with linear dependencies and objects, or we discuss their linearizations. In statistics, many methods belong to the linear models, too. We consider a quite general scheme of this type.



Consider a random vector  $Y = (Y_1, \dots, Y_n)^T$  and suppose

$$Y = X \cdot \beta + \sigma Z,$$

where  $X = (x_{ij})$  is a constant real-valued matrix with  $n$  rows and  $k < n$  columns, whose rank is  $k$ ,  $\beta$  is an unknown constant vector of  $k$  parameters of the model,  $Z$  is a random vector whose  $n$  components have distribution  $N(0, 1)$ , and  $\sigma > 0$  is an unknown positive parameter of the model. This is a *linear model* with full rank.

In practice, the variables  $x_{ij}$  are often known. The problem is to estimate or predict the value of  $Y$ . For instance,  $x_{ij}$  can express the grade in maths of the  $i$ -th student in the  $j$ -th semester ( $j = 1, 2, 3$ ), and we want to know how this student will fare in the fourth semester. For this purpose, the vector  $\beta$  needs to be known. This can be estimated using complete observations, that is, from the knowledge of  $Y$  (from the results of past years, for example).

In order to estimate the vector  $\beta$ , the *least squares method* can often be used. This means looking for the estimate  $b \in \mathbb{R}^k$  for which the vector  $\hat{Y} = Xb$  minimizes the squared length of the vector  $Y - X\beta$ .

This is a simple problem from linear algebra, looking for the orthogonal projection of the vector  $Y$  onto the subspace  $\text{span } X \subset \mathbb{R}^n$  generated by the columns of the matrix  $X$ . This is minimizing the function

$$\|Y - X\beta\|^2 = \sum_{i=1}^n \left( Y_i - \sum_{j=1}^k x_{ij} \beta_j \right)^2.$$

Choose an arbitrary orthonormal basis of the vector subspace  $\text{span } X$  and write it into columns of the matrix  $P$ . For any choice of basis, the orthogonal projection is realized as multiplication by the matrix  $PP^T$ . In the subspace  $\text{span } X$ , the mapping given by this matrix is the identity. That is,

$$\hat{Y} = PP^T Y = PP^T (X\beta + \sigma Z) = X\beta + \sigma PP^T Z.$$

The matrix  $PP^T$  is positive-semidefinite. Extend the basis consisting of the columns of  $P$  to an orthonormal basis of the whole space  $\mathbb{R}^n$ . In other words, create a matrix  $Q = (P \ R)$  by writing the newly added basis vectors into the matrix  $R$  with  $n - k$  columns and  $n$  rows. Denote by  $V = P^T Z$  and  $U = R^T Z$  the random vectors with  $k$  and  $n - k$  components, respectively. They are orthogonal, and their sum in  $\mathbb{R}^n$  is the vector  $(V^T \ U^T)^T = Q^T Z$ .



Clearly (see subsection 10.2.46), both vectors  $V$  and  $U$  have multivariate normal distribution with zero expected value and identity covariance matrix. The random vector  $Y$  is decomposed to the sum of a constant vector  $X\beta$  and two orthogonal projections

$$Y = X\beta + \sigma PV + \sigma RU,$$

and the desired orthogonal projection is the sum of the first and second summands. In subsection 10.2.46, the distribution of such random vectors is also derived.

The size of  $\|Y - \hat{Y}\|^2$  is called the *residual sum of squares*, sometimes denoted by *RSS*. Also, the *residual variance* is defined as

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Recall that  $\hat{Y} = Xb$  and that  $X^T X$  is invertible as the full rank of  $X$  is assumed. Thus  $b = (X^T X)^{-1} X^T \hat{Y}$  can be computed. At the same time,  $X^T (Y - \hat{Y}) = \sigma X^T (RU) = 0$ , since the columns of  $X$  and  $R$  are mutually orthogonal. Therefore,

$$(1) \quad b = (X^T X)^{-1} X^T Y.$$

The chosen matrix  $P$  can be used with advantage. Since its columns generate the same subspace as the columns of  $X$ , there is a square matrix  $T$  such that  $X = PT$  (its columns are the coefficients of linear combinations which are expressed by the columns of  $X$  in the basis of  $P$ ). Substitute (using the fact that  $P^T P$  is the identity matrix and  $T$  is invertible):

$$\begin{aligned} b &= (T^T P^T P T)^{-1} T^T P^T Y = \\ &= T^{-1} (T^T)^{-1} T^T P^T (PT\beta + \sigma Z) = \\ &= \beta + \sigma T^{-1} V. \end{aligned}$$

Thus is proved the main properties of the linear model:

**Theorem.** Consider a linear model  $Y = X\beta + \sigma Z$ .

(1) For the estimate of  $\hat{Y}$ ,

$$\hat{Y} = X\beta + \sigma PV, \quad \hat{Y} \sim N(X\beta, \sigma^2 P P^T).$$

(2) The residual sum of squares and the normed square of the residue size have distributions:

$$Y - \hat{Y} \sim N(0, \sigma^2 R R^T), \quad \|Y - \hat{Y}\|^2 / \sigma^2 \sim \chi_{n-k}^2.$$

(3) The random variable  $b = \beta + \sigma T^{-1} V$  has distribution

$$b \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

(4) The residual variance satisfies  $(n - k)S^2 / \sigma^2 \sim \chi_{n-k}^2$ .

(5) The expected value of the residual variance is  $E S^2 = \sigma^2$ .

(6) The variables  $b$  and  $S^2$  are independent.

**PROOF.** Both the shape and distribution of  $\hat{Y}$  are determined. It is clear that  $Y - \hat{Y} = \sigma RU$ , which verifies the second proposition. Further,

$$\|Y - \hat{Y}\|^2 / \sigma^2 = \|RU\|^2 = \|U\|^2,$$

where the last equality follows from the fact that in the construction,  $U$  is the vector of coordinates of the projection  $Z$  onto the complement of  $\text{span } X$ , and  $RU$  is this projection. The size of a vector is exactly the sum of squares of its coordinates in any orthonormal basis.

Therefore, the random vector  $\|Y - \hat{Y}\|^2/\sigma^2$  is the sum of  $(n-k)$  squares of random variables with distribution  $N(0, 1)$ , so it is the distribution  $\chi_{n-k}^2$ , which proves the rest of (2).

The next proposition follows directly from the definitions and calculations. It suffices to estimate the covariance matrix for  $b$ . From the general properties, it should be the matrix  $T^{-1}(T^T)^{-1}$ . This is the same as  $(X^T X)^{-1} = ((PT)^T(PT))^{-1}$ .

The proposition (4) is a reformulation of the information in (2). The next proposition follows from the fact that the expected value of the  $\chi^2$  distribution equals the number of degrees of freedom.

Finally, independence of the variables  $b$  and  $S$  is a consequence of the fact that the former variable is a function of the vector  $V$ , while the latter one is a function of the vector  $U$ . These vectors are independent since they are two complementary parts from an orthogonal transformation of the vector  $Z$ .  $\square$

In practice, the hypothesis whether fewer parameters are sufficient to estimate the expected value is sometimes tested. A random vector  $Y$  is said to satisfy a *submodel* if and only if both  $Y = X\beta + \sigma Z$  and



$$Y = X^0 \beta^0 + \sigma Z,$$

where  $X^0$  has only  $q < k$  columns. It is assumed that the columns of  $X^0$  generate a subspace in  $\text{span } X$ , i.e., all are linear combinations of the columns of  $X$ .

Repeat the above construction, choosing the matrix  $P$  so that the first  $q$  vectors of  $P$  generate  $\text{span } X^0$ . The matrix  $P$  is then of the form  $(P^0 P^1)$ , and the vector  $V$  decomposes similarly:

$$V = \begin{pmatrix} V^0 \\ V^1 \end{pmatrix} = \begin{pmatrix} (P^0)^T Z \\ (P^1)^T Z \end{pmatrix}.$$

This yields a finer decomposition of the vectors and their sizes and the corresponding residues:

$$\hat{Y}^0 = P^0(P^0)^T Y = X^0 \beta^0 + \sigma P^0 V^0$$

$$Y - \hat{Y}^0 = \sigma P^1 V^1 + \sigma R U$$

$$\|Y - \hat{Y}^0\|^2 = \sigma^2 \|V^1\|^2 + \sigma^2 \|U\|^2$$

$$(\text{RSS}^0 - \text{RSS})/\sigma^2 = \|V^1\|^2.$$

Therefore, the normed difference of the residues has distribution  $\chi_{k-q}^2$ . It follows immediately that the statistic  $F$  given as the relative difference of the residues has Fisher-Snedecor distribution:

$$F = \frac{(\text{RSS}^0 - \text{RSS})/(k-q)}{\text{RSS}/(n-k)} \sim F_{k-q, n-k}.$$

In practice, the parameter  $\sigma$  is seldom known, and so the estimate  $S^2$  is used. Instead of the individual components  $b_j \sim N(\beta_j, \sigma^2 c_{jj})$  of the random vector  $b$ , where  $c_{jj}$  are the diagonal entries of the matrix  $C = (X^T X)^{-1}$ , work with the statistics

$$T_j = \frac{b_j - \beta_j}{S\sqrt{c_{jj}}} \sim t_{n-k}.$$

Of course, these variables need not be independent.

If the full rank of the matrix  $X$  is not assumed, a pseudoinverse matrix can be used instead of  $C = (X^T X)^{-1}$ .

**10.3.12. Examples of tests.** As an illustration, we mention some examples of application of linear models in the simplest types of tests. The most trivial case is when there is only one sample. Here the test is whether or not the only parameter  $\beta$  equals a given value  $\beta_0$ .

For this case, choose the matrix  $X$  as a single column consisting of ones. Then, the expression

$$Y = X\beta + \sigma Z$$

indicates that the individual components in  $Y$  are independent variables  $Y_i \sim N(\beta, \sigma^2)$ . It is a random sample of size  $n$  from the normal distribution. In general, the estimate

$$b = (X^T X)^{-1} X^T Y = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$$

$$S^2 = \frac{1}{n-1} \|Y - X\bar{Y}\|^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

which respectively is exactly the sample mean and variance used before.

In this context, the statistic

$$T = \frac{\bar{Y} - \beta_0}{S} \sqrt{n}$$

may also be of interest.

Testing the hypothesis  $\beta = \beta_0$  is called the *one-sample t-test*. The hypothesis is rejected at level  $\alpha$  if  $|T| \geq t_{n-1}(\alpha)$ .

There is another simple application of the general model, which is called the *paired t-test*. It is appropriate for cases when pairs of random vectors  $W_1 = (W_{i1})$  and  $W_2 = (W_{i2})$  are tested. The differences  $Y_i = W_{i1} - W_{i2}$  of their components have distribution  $N(\beta, \sigma^2)$ . In addition, the variables  $Y_i$  need to be independent (which does not mean that the individual pairs  $W_{i1}$  and  $W_{i2}$  have to be independent!). In the context of our illustrative example from 10.3.4, we can imagine the assessment of two lecturers by the same student.

Test the hypothesis that for every  $i$ ,  $E W_{i1} = E W_{i2}$ . Thus, use the statistic

$$T = \frac{\bar{W}_1 - \bar{W}_2}{S} \sqrt{n}.$$

Finally, we consider an example with more parameters. It is a classical case of the *regression line*.

Assume that the variables  $Y_i, i = 1, \dots, n$  have distribution  $N(\beta_0 + \beta_1 x_i, \sigma^2)$ , where  $x_i$  are given constant. Examine the best approximation

$$Y_i = b_0 + b_1 x_i,$$

and the matrix  $X$  of the corresponding linear model is

$$X^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}.$$

Substitute into general formulae, and compute the estimate

$$\begin{aligned} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} &= \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} = \\ &= \begin{pmatrix} \sum_{i=1}^n (x_i - \bar{x})^2 \\ \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}. \end{aligned}$$

It follows that

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Finally, compute  $b_0 = \bar{Y} - b_1 \bar{x}$ . From the calculations,

$$\text{var } b_1 = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2.$$

In order to test the hypothesis whether the expected value of a variable  $Y$  does not depend on  $x$ , that is, whether or not  $H_0$  is of the form  $\beta_1 = 0$ , use the statistic

$$T = \frac{b_1}{S} \left( \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{1/2} \sim t_{n-2}.$$

The statistical analysis of multiple regression is similar. There are several sets of values  $x_{ij}$  to evaluate the statistical relevance of the approximation

$$Y_i = b_0 + b_1 x_{1i} + \dots + b_k x_{ki}.$$

The individual statistics  $T_j$  allow for a t-test of dependence of the regression on the individual parameters. Software packages often provide a parameter which expresses how well the values  $Y_i$  are approximated. It is called the coefficient of determination:

$$R^2 = 1 - \frac{\text{RSS}}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

**10.3.13.** In practice, problems are often met where the distributions of the statistical data sets are either completely unknown or errors are assumed in the model, together with deviations with non-zero expected value and a non-normal distribution. In these cases, application of classical frequentist statistics is very hard or even totally impossible.



There are approaches which work directly with the sample set. Then derive statistics of point or interval estimates or probability calculations about the above, including the evaluation of standard errors.

One of the pioneering articles of this topic is the brief work of Bradley Efron of Stanford University, published in

1981: *Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods*<sup>4</sup>. The keywords of this article are: Balanced repeated replications; Bootstrap; Delta method; Half-sampling; Jackknife; Infinitesimal jackknife; Influence function.

The procedure used in the *bootstrap* method uses software resources, created from a given data sample, and new data samples of the same size (with replacement). The desired statistics (sample mean, variance, etc.) is then examined for each of them. After a great number of executions of this procedure, a data sample is obtained which is considered a relevant approximation of the probability distribution of the examined statistic. The characteristics of this data set is considered a good approximation of the characteristics of the examined statistics for point or interval estimates, analysis of variance, etc. There is not enough space for a more detailed analysis of these techniques, which is the foundations of non-parametric methods in contemporary software statistical tools.

---

<sup>4</sup>Biometrika (1981), 68, 3, pp. 589-99

## Number theory

*God created the integers, all else is the work of man.*  
*Leopold Kronecker*



### A. Basic properties of divisibility

Let us recall the basic properties of divisibility, whose proof follows directly from the definition: the integer 0 is divisible by every integer; the only integer that is divisible by 0 is 0; every integer  $a$  satisfies  $a \mid a$ ; every triple of integers  $a, b, c$  satisfies the following four implications:



$$\begin{aligned} a \mid b \wedge b \mid c &\implies a \mid c, \\ a \mid b \wedge a \mid c &\implies (a \mid b + c) \wedge (a \mid b - c), \\ c \neq 0 &\implies (a \mid b \iff ac \mid bc), \\ a \mid b \wedge (b > 0) &\implies a \leq b. \end{aligned}$$

The mere knowledge of these basic rules allows us to solve many problems.

**11.A.1.** Determine the natural numbers  $n$  for which the integer  $n^3 + 1$  is divisible by the integer  $n - 1$ .

**Solution.** We have  $n^3 - 1 = (n - 1)(n^2 + n + 1)$ , so the integer  $n^3 - 1$  is divisible by the integer  $n - 1$  for any  $n$ . If  $n - 1$  is to divide  $n^3 + 1$  as well, it must also divide the difference  $(n^3 + 1) - (n^3 - 1) = 2$  (see the second property of divisibility). Since  $n \in \mathbb{N}$ , we have  $n - 1 \geq 0$ . Now,  $n - 1 \mid 2$  implies that  $n - 1 = 1$  or  $n - 1 = 2$ , whence  $n = 2$

In this chapter, we will deal with problems concerning integers, which include mainly divisibility and solving equations whose domain will be the set of integers (or natural numbers). Notice that in this chapter, unlike in the other parts of this book, we will not consider zero to be a natural number, as is usual in this field of mathematics. Although the natural numbers and the integers are, from a certain point of view, the simplest mathematical structure, examination of their properties yielded a good deal of tough problems for generations of mathematicians. These are often problems which can be formulated quite easily, yet many of them remain unsolved so far. Let us mention some of the most popular of them:

- *twin primes* – the problem is to decide whether there are infinitely many primes  $p$  such that  $p + 2$  is also a prime,<sup>1</sup>
- *Sophie Germain primes* – the problem is to decide whether there are infinitely many primes  $p$  such that  $2p + 1$  is also a prime,
- *existence of an odd perfect integer* – i.e., the sum of whose divisors equals twice the integer, e.g., 6 or 28,
- *Goldbach's conjecture* – the problem is to decide whether every even integer greater than 2 can be expressed as the sum of two primes.

A jewel among the problems of number theory is the *Fermat's Last Theorem* – the problem is to decide whether there are natural numbers  $n, x, y, z$  such that  $n > 2$  and  $x^n + y^n = z^n$ ; Pierre de Fermat formulated this problem as early as in 1637; much effort of many generations was put to this question, and it was solved (using results of various fields of mathematics) by Andrew Wiles in 1995.

### 1. Fundamental concepts

**11.1.1. Divisibility.** Recall that we say that an integer  $a$  divides an integer  $b$  (or that  $b$  is *divisible* by  $a$ , also that  $b$  is a *multiple* of  $a$ ) if and only if there exists an integer  $c$  satisfying  $a \cdot c = b$ . We write this as  $a \mid b$ . The concept of divisibility can be considered much more generally, as we shall see in 12.3.5.



<sup>1</sup>In 2013, Yitang Zhang published a proof of a promising proposition: for some  $n < 7 \cdot 10^7$ , there are infinitely many pairs of primes which differ by  $n$ . See Y. Zhang, *Bounded gaps between primes*, *Annals of Mathematics*, 2013. Although the bound was improved to  $N = 246$  already one year later by James Maynard and Terence Tao, the problem is still open.

or  $n = 3$ . The wanted property is thus possessed only by the natural numbers 2 and 3.  $\square$

**11.A.2.** Prove that for any  $a \in \mathbb{Z}$ , the following holds:

- i)  $a^2$  leaves remainder 0 or 1 when divided by 4;
- ii)  $a^2$  leaves remainder 0, 1, or 4 when divided by 8;
- iii)  $a^4$  leaves remainder 0 or 1 when divided by 16.

**Solution.**

- It follows from the Euclidean division theorem that every integer  $a$  can be written uniquely in either the form  $a = 2k$  or  $a = 2k + 1$ . Squaring this leads to  $a^2 = 4k^2$  or  $a^2 = 4(k^2 + k) + 1$ , which is what we wanted to prove.
- Making use of the above result, we immediately obtain the statement for the (even) integers of the form  $a = 2k$ . Back then, we arrived at  $a^2 = 4k(k + 1) + 1$  for odd integers  $a$ ; we get the proposition easily if we realize that  $k(k + 1)$  is surely even.
- Again, we utilize the result of the previous parts, i.e.  $a^2 = 4\ell$  or  $a^2 = 8\ell + 1$ . Squaring these equalities once again, we get  $a^4 = (a^2)^2 = 16\ell^2$  for  $a$  even, and  $a^4 = (a^2)^2 = (8\ell + 1)^2 = 64\ell^2 + 16\ell + 1 = 16(4\ell^2 + \ell) + 1$  for  $a$  odd.  $\square$

**11.A.3.** Prove that if integers  $a, b \in \mathbb{Z}$  leave remainder 1 when divided by an  $m \in \mathbb{N}$ , then so does their product  $ab$ .

**Solution.** By the Euclidean division theorem, there are  $s, t \in \mathbb{Z}$  such that  $a = sm + 1, b = tm + 1$ . Multiplying these equalities leads to the expression

$$ab = (sm + 1)(tm + 1) = (stm + s + t)m + 1,$$

where  $stm + s + t$  is the quotient, so the remainder of  $ab$  when divided by  $m$  is equal to 1.  $\square$

*It follows from the Euclidean division theorem that the greatest common divisor  $(a, b)$  of any pair of integers  $a, b$  exists, is unique, and can be computed efficiently by the Euclidean algorithm. At the same time, the coefficients in Bézout's identity, i.e., integers  $k, l$  such that  $ka + lb = (a, b)$ , can be determined this way. It can also be easily proved straight from the properties of divisibility that integer linear combinations of integers  $a, b$  are exactly the multiples of their greatest common divisor.*

**11.A.4.** Find the greatest common divisor of the integers  $a = 10175, b = 2277$  and determine the corresponding coefficients in Bézout's identity.

**Solution.** We will invoke the Euclidean algorithm:

$$\begin{aligned} 10175 &= 4 \cdot 2277 + 1067, \\ 2277 &= 2 \cdot 1067 + 143, \\ 1067 &= 7 \cdot 143 + 66, \\ 143 &= 2 \cdot 66 + 11, \\ 66 &= 6 \cdot 11 + 0. \end{aligned}$$

We will often take advantage of one of the most important properties of the integers, the unique *Euclidean division* (i.e., division with remainder).

UNIQUE DIVISION WITH REMAINDER

**Theorem.** For any integers  $a \in \mathbb{Z}, m \in \mathbb{N}$ , there exists a unique pair of integers  $q \in \mathbb{Z}, r \in \{0, 1, \dots, m - 1\}$  satisfying  $a = qm + r$ .

**PROOF.** First, we will prove the existence of the integers  $q, r$ . Let us fix a natural number  $m$  and prove the statement for any  $a \in \mathbb{Z}$ . Assume first that  $a$  is non-negative and prove the existence of the integers  $q, r$  by induction on  $a$ :

If  $0 \leq a < m$ , we can choose  $q = 0, r = a$ , and the equality  $a = qm + r$  holds trivially.

Next, suppose that  $a \geq m$  and the existence of the integers  $q, r$  has been proved for all  $a' \in \{0, 1, 2, \dots, a - 1\}$ . In particular, for  $a' = a - m \geq 0$ , there are  $q', r'$  such that  $a' = q'm + r'$  and  $r' \in \{0, 1, \dots, m - 1\}$ .

Therefore, if we select  $q = q' + 1, r = r'$ , we obtain  $a = a' + m = (q' + 1)m + r' = qm + r$ , which is what we wanted to prove.

Now, if  $a$  is negative, then we have proved that for the positive integer  $-a$ , there are  $q' \in \mathbb{Z}, r' \in \{0, 1, \dots, m - 1\}$  such that  $-a = q'm + r'$ . If  $r' = 0$ , we set  $r = 0, q = -q'$ ; otherwise (i.e.,  $r' > 0$ ), we put  $r = m - r', q = -q' - 1$ .

In either case, we get  $a = q \cdot m + r$ . Therefore, the integers  $q, r$  with the wanted properties exist for every  $a \in \mathbb{Z}, m \in \mathbb{N}$ .

Finally, we will prove the uniqueness. Suppose that there are integers  $q_1, q_2 \in \mathbb{Z}$  and  $r_1, r_2 \in \{0, 1, \dots, m - 1\}$  which satisfy  $a = q_1m + r_1 = q_2m + r_2$ . Simple rearrangement yields  $r_1 - r_2 = (q_2 - q_1)m$ , so  $m \mid r_1 - r_2$ . However, we have  $0 \leq r_1 < m$  and  $0 \leq r_2 < m$ , whence it follows that  $-m < r_1 - r_2 < m$ .

Therefore,  $r_1 - r_2 = 0$ , and  $(q_2 - q_1)m = 0$ , hence  $q_1 = q_2, r_1 = r_2$ .  $\square$

The integers  $q$  and  $r$  from the theorem are respectively called the *quotient* and *remainder* of the division of  $a$  by  $m$  with remainder. The choice of this terminology seems more intuitive if we rearrange the equality  $a = mq + r$  into the form

$$\frac{a}{m} = q + \frac{r}{m}, \quad \text{where } 0 \leq \frac{r}{m} < 1.$$

**11.1.2. Greatest common divisor.** One of the most needed tools of computational number theory is the algorithm for computing the greatest common divisor. Since it is a relatively fast procedure, as we are going to show, it is used very often in modern algorithms as well.







$$\begin{aligned} 2^7 - 1 &= k(2^{49} - 1) + \ell(2^{35} - 1): \\ 2^7 - 1 &= (2^{35} - 1) - (2^{21} + 2^7)(2^{14} - 1) \\ &= (2^{35} - 1) - (2^{21} + 2^7)((2^{49} - 1) - 2^{14}(2^{35} - 1)) \\ &= (2^{35} + 2^{21} + 1)(2^{35} - 1) - (2^{21} + 2^7)(2^{49} - 1). \end{aligned}$$

Therefore,  $k = -(2^{21} + 2^7)$ ,  $\ell = 2^{35} + 2^{21} + 1$ . Let us bear in mind that these coefficients are never determined uniquely.  $\square$

**11.A.7.** Now, let us try to generalize the result of the previous exercise, i.e., prove that it holds for any  $a, m, n \in \mathbb{N}$ ,  $a \neq 1$ , that  $(a^m - 1, a^n - 1) = a^{(m,n)} - 1$ .

**Solution.** This statement follows easily from the fact that any pair of natural numbers  $k, \ell$  satisfies  $a^k - 1 \mid a^\ell - 1$  if and only if  $k \mid \ell$ . This can be proved by dividing the integer  $\ell$  by the integer  $k$  with remainder, i.e., we set  $\ell = kq + r$ , where  $q, r \in \mathbb{N}_0, r < k$ , and consider that

$$a^{kq+r} - 1 = (a^k - 1)(a^{k(q-1)+r} + a^{k(q-2)+r} + \dots + a^r) + a^r - 1$$

is the division of the integer  $a^{kq+r} - 1$  by the integer  $a^k - 1$  with remainder (apparently, we have  $a^r - 1 < a^k - 1$ ). Hence we can easily see that the remainder  $r$  is zero if and only if the remainder  $a^r - 1$  is zero, which is what we wanted to prove.  $\square$

**11.A.8.** Prove that for all  $n \in \mathbb{N}$ ,  $25 \mid 4^{2n+1} - 10n - 4$ .

**Solution.** This statement can be proved in several ways (the easiest one is in terms of congruences, which will be introduced a bit later). Here, we will prove it by induction and then as a consequence of the binomial theorem.

The proposition is clearly true for  $n = 0$  (even though the problem does not ask about the situation for  $n = 0$ , we can surely prove the desired property for the set  $n \in \mathbb{N}_0$ , thereby simplifying the first step of the induction). As for the second step: assuming  $25 \mid 4^{2n-1} - 10(n-1) - 4$ , then we also have  $25 \mid 16(4^{2n-1} - 10(n-1) - 4) = 4^{2n+1} - 10n - 4 - 150n + 100$ , whence it easily follows that the desired proposition  $25 \mid 4^{2n+1} - 10n - 4$  holds.

The second proof uses the binomial theorem. By that,

$$(5 - 1)^{2n+1} = \sum_{k=0}^{2n+1} \binom{2n+1}{k} 5^{2n+1-k} (-1)^k,$$

where all of the terms of the sum, except the last two, are apparently multiples of 25, i.e., the only part of the sum which is not divisible by 25 is  $\binom{2n+1}{2n} 5^1 (-1)^{2n} + 5^0 (-1)^{2n+1}$ . In other words,  $4^{2n+1}$  leaves the same remainder when divided by 25 as the integer  $10n + 4$ , which is equivalent to what we were to prove.  $\square$

## B. Prime numbers

*Euclid's theorem 11.2.1 is a fundamental property of primes as it can be seen in the following example.*

On the other hand, any common divisor of the given integers  $a_1, a_2$  divides the integer  $a_3 = a_1 - q_1 a_2$  as well, hence it also divides  $a_4 = a_2 - q_2 a_3, a_5, \dots$ , and especially  $a_{k-1} = a_{k-3} - q_{k-3} a_{k-2}$ . We have thus proved that  $a_{k-1}$  is the greatest common divisor of the integers  $a_1, a_2$ .  $\square$

It follows from the previous statement and the fact that  $(a, b) = (a, -b) = (-a, b) = (-a, -b)$  holds for any  $a, b \in \mathbb{Z}$ , that every pair of integers has a greatest common divisor.

**11.1.3. Bézout's theorem.** The Euclidean algorithm provides another interesting statement, which is often used.

**Theorem (Bézout).** For every pair of integers  $a, b$ , there exist integers  $k, l$  such that  $(a, b) = ka + lb$ .

**PROOF.** Surely it suffices to prove the theorem for  $a, b \in \mathbb{N}$ . Let us notice that if it is possible to express integers  $r, s$  in the form  $r = r_1 a + r_2 b, s = s_1 a + s_2 b$ , where  $r_1, r_2, s_1, s_2 \in \mathbb{Z}$ , then we can also express

$$r + s = (r_1 + s_1)a + (r_2 + s_2)b$$

in this way as well as

$$c \cdot r = (c \cdot r_1)a + (c \cdot r_2)b$$

for any  $c \in \mathbb{Z}$ , and thus also any integer linear combination of the numbers  $r$  and  $s$  arising in the process of Euclidean algorithm. It follows from the Euclidean algorithm (for  $a_1 = a, a_2 = b$ ) that we can also thus express  $a_3 = a_1 - q_1 a_2, a_4 = a_2 - q_2 a_3, \dots$ , hence the integer  $a_{k-1} = a_{k-3} - q_{k-3} a_{k-2}$ , too, which is  $(a_1, a_2)$ .

Let us emphasize that the wanted numbers  $k, l$  are *not* determined uniquely.  $\square$

The Euclidean algorithm and Bézout's identity are the fundamental results of elementary number theory and form one of the pillars of the algorithms used in algebra and number theory.

**11.1.4. Least common multiple.** We have ignored the properties of the least common multiple so far. However, we will see that thanks to the following proposition, they can be derived from the properties of the greatest common divisor.



**Lemma.** For every pair of integers  $a, b$ , their least common multiple  $[a, b]$  exists and it holds that  $(a, b) \cdot [a, b] = |a \cdot b|$ .

**PROOF.** The statement is trivially true if either of the integers  $a, b$  is zero. Furthermore, we can assume that both these (non-zero from now on) integers are positive since their signs do not take any effect in the formula in question. We are going to show that  $q = a \cdot b / (a, b)$  is the least common multiple of the integers  $a, b$ , which will finish the proof.

Since  $(a, b)$  is a common divisor of  $a, b$ , both  $a/(a, b)$  and  $b/(a, b)$  are integers, hence

$$q = \frac{ab}{(a, b)} = \frac{a}{(a, b)} \cdot b = \frac{b}{(a, b)} \cdot a$$

is a common multiple of  $a, b$ . By Bézout's identity, there are integers  $k, l$  such that  $(a, b) = ka + lb$ . Let us suppose

**11.B.1.**

- i) Prove that if natural numbers  $m, n$  are coprime, then so are

$$m^2 + mn + n^2 \quad \text{and} \quad m^2 - mn + n^2.$$

- ii) Prove that if odd natural numbers  $m, n$  are coprime, then so are

$$m + 2n \quad \text{and} \quad m^2 + 4n^2.$$

**Solution.**

- i) To reach a contradiction, suppose that there is a prime  $p$  which divides both of the integers  $m^2 + mn + n^2$  and  $m^2 - mn + n^2$ . Then, it divides their difference  $2mn$  as well, whence  $p = 2$  or  $p$  divides one of the integers  $m, n$ . If  $p = 2$ , then  $m^2 + mn + n^2$  is even, so the integers  $m$  and  $n$  must be even as well, which contradicts that they are coprime. If  $p$  divides  $m$  as well as  $m^2 + mn + n^2$ , then it also divides  $n^2$ , whence, by Euclid's theorem (11.2.1), it divides  $n$  as well. However, this also contradicts that  $m, n$  are coprime. The case of  $p \mid n$  is analogous.
- ii) Just like in the above exercise, let us suppose that there is a prime  $p$  which divides  $m + 2n$  as well as  $m^2 + 4n^2$ . Then, it must also divide  $(m^2 + 4n^2) - (m + 2n)(m - 2n) = 8n^2$ , and since  $p \neq 2$  (if  $m + 2n$  were even, then so would  $m$  be), we necessarily have  $p \mid n$ . However, since  $p$  divides  $m + 2n$  as well, we get  $p \mid m$ , which is a contradiction.

**11.B.2.** Prove that any  $n \in \mathbb{N}, n > 1$ , is prime iff  $n$  is not divisible by any prime  $p \leq \sqrt{n}$ .

**Solution.** For  $n$  composite we have  $n = ab$  with appropriate  $a, b > 1$ . If we admitted  $a, b > \sqrt{n}$  then we would have  $n = ab > \sqrt{n} \cdot \sqrt{n} = n$ . Therefore  $n$  has a divisor (and thus also a prime divisor) not greater than  $\sqrt{n}$ .  $\square$

The theoretical part contains Euclid's proof of the infinitude of primes and deals in detail with the distribution of primes in the set of natural numbers (in some cases, however, we were forced to leave the mentioned theorems unproved). Now, we will give several exercises on this topic.

**11.B.3.** For any natural number  $n \geq 3$ , there is at least one prime between the integers  $n$  and  $n!$ .



**Solution.** Let  $p$  denote an arbitrary prime dividing the integer  $n! - 1$  (by the Fundamental theorem of arithmetic (11.2.2), there is such a prime since  $n! - 1 > 1$ ). If we had  $p \leq n$ , then  $p$  would have to divide  $n!$  as well, so it could not divide  $n! - 1$ . Therefore,  $n < p$ . Since  $p \mid (n! - 1)$ , we have  $p \leq n! - 1$ , hence  $p < n!$ . Our prime  $p$  thus satisfies the conditions of the problem.  $\square$

The result of this exercise also implies the infinitude of primes (it suffices to consider the sequence  $a_0 = 3, a_{n+1} = a_n!$  for  $n \in \mathbb{N}$ ). However, this statement is very weak (compared to reality) since the constructed sequence contains only a "tiny" subset of the primes.

On the other hand, we are able to construct an arbitrarily long sequence of consecutive composite numbers, as shown by the following exercise.

that  $n \in \mathbb{Z}$  is an arbitrary common multiple of the integers  $a, b$ . We want to show that it is divisible by  $q$ . We thus have  $n/a, n/b \in \mathbb{Z}$ , hence the number

$$\frac{n}{b} \cdot k + \frac{n}{a} \cdot l = \frac{n(ka + lb)}{ab} = \frac{n(a, b)}{ab} = \frac{n}{q}$$

is an integer as well. However, this means that  $q \mid n$ , which is what we wanted to prove.  $\square$

**11.1.5. Coprime integers.** Analogously to the case of two integers, we can also define the greatest common divisor and least common multiple of more than two integers, and it can be easily proved that



$$(a_1, \dots, a_n) = ((a_1, \dots, a_{n-1}), a_n),$$

$$[a_1, \dots, a_n] = [[a_1, \dots, a_{n-1}], a_n].$$

Integers  $a_1, a_2, \dots, a_n \in \mathbb{Z}$  are said to be *coprime* (also relatively prime) if and only if  $(a_1, a_2, \dots, a_n) = 1$ . They are said to be *pairwise coprime* (pairwise relatively prime) if and only if we have  $(a_i, a_j) = 1$  for every pair of indices  $i, j$  satisfying  $1 \leq i < j \leq n$ .

**Remark.** Let us realize that the concepts *coprime* and *pairwise coprime* are different. For example, we have  $(6, 10, 15) = 1$ ; however, any two of the three integers 6, 10, 15 are not coprime.

**Lemma.** For any natural numbers  $a, b, c$  we have

- (1)  $(ac, bc) = (a, b) \cdot c$ ;
- (2) if  $a \mid bc$  and  $(a, b) = 1$ , then  $a \mid c$ ;
- (3)  $d = (a, b)$  if and only if there are  $k, l \in \mathbb{N}$  such that  $a = dk, b = dl$ , and  $(k, l) = 1$ .

**PROOF.** (1) Since  $(a, b)$  is a common divisor of the integers  $a, b$ ,  $(a, b) \cdot c$  is a common divisor of the integers  $ac, bc$ , hence  $(a, b) \cdot c \mid (ac, bc)$ . From Bézout's identity, we obtain  $k, l \in \mathbb{Z}$  such that  $(a, b) = ka + lb$ . Since  $(ac, bc)$  is a common divisor of the integers  $ac, bc$ , it divides the integer  $kac + lbc = (a, b) \cdot c$  as well. We have thus proved that  $(a, b) \cdot c$  and  $(ac, bc)$  are natural numbers which divide each other, hence they are equal.

(2) Let us suppose that  $(a, b) = 1$  and  $a \mid bc$ . From Bézout's identity again, we get  $k, l \in \mathbb{Z}$  such that  $ka + lb = 1$ , whence it follows that  $c = c(ka + lb) = kca + lbc$ . Since  $a \mid bc$ , it follows that  $c$  as well is a multiple of  $a$ .

(3) Let  $d = (a, b)$ , then there are  $q_1, q_2 \in \mathbb{N}$  such that  $a = dq_1, b = dq_2$ . Then, by part (1), we have  $d = (a, b) = (dq_1, dq_2) = d \cdot (q_1, q_2)$ , so  $(q_1, q_2) = 1$ . On the other hand, if  $a = dq_1, b = dq_2$ , and  $(q_1, q_2) = 1$ , then  $(a, b) = (dq_1, dq_2) = d(q_1, q_2) = d \cdot 1 = d$  (again invoking part (1) of this lemma).  $\square$

**2. Primes**

The concept of a prime is one of the most important in elementary number theory. Its importance is given mainly by the unique factorization theorem, which is a strong as well as efficient tool for solving miscellaneous problems from number theory.



**11.B.4.** Prove that for any natural number  $n$ , there exist  $n$  consecutive natural numbers none of which is prime.

**Solution.** Let us examine the integers  $(n + 1)! + 2$ ,  $(n + 1)! + 3, \dots, (n + 1)! + (n + 1)$ . For any  $k \in \{2, 3, \dots, n+1\}$ , we have  $k \mid (n+1)!$ , so  $k \mid (n+1)!+k$  as well, thus  $(n + 1)! + k$  cannot be a prime. Therefore, there are no primes among these  $n$  natural numbers.  $\square$

**Practical notes.** As we will show, it is very complicated to decide for certain whether a given large integer is a prime (on the other hand, for most composite numbers, it is really easy to prove that they are indeed composite – see part 11.6.4). Nevertheless, Indian mathematicians<sup>1</sup> managed to prove in 2002 that there is an algorithm running in polynomial time with respect to the input (i.e., the number of digits of the integer in question) which decides whether the integer is a prime.



We are unable to produce such an algorithm for prime factorization (and it is widely believed that it is impossible although no one has been able to prove this so far). The fastest generally applicable factorization algorithm, the so-called *general number field sieve*, runs in sub-exponential time  $O\left(e^{1.9(\log N)^{1/3}(\log \log N)^{2/3}}\right)$ .

In 1994, Peter Shor invented an algorithm which factors an integer  $N$  in cubic time (i.e., runs it  $O(\log^3 N)$ ) on a quantum computer. However, this algorithm requires computers with sufficient number of quantum bits. We can see how difficult this is from the fact that in 2001, IBM managed to use a quantum computer to factor the integer 15, and in 2012, another record was achieved by factoring the integer 143 (in fact using other approach, so-called adiabatic quantum computation).

We can find more evidence about the difficulty of this problem in the call made in 1991 by RSA Security.<sup>2</sup> If anyone manages to factor the integers labeled by the number of digits they have as RSA-100, ..., RSA-704, RSA-768, ..., RSA-2048, they could receive respectively 1,000, ..., 30,000, 50,000, ..., 200,000 dollars (the integer RSA-100 was factored by Arjen Lenstra that very year; the integer RSA-704 was factored in 2012; many others have not been factored yet).

*Thanks to the unique factorization theorem, we are able to (provided we know this factorization) easily answer the following questions concerning the number or sum of the divisors of a given integer. Just that easily, we can get the (intuitively well-known) procedure of computation of the greatest common divisor of two integers from their prime factorizations.*

<sup>1</sup>M. Agrawal, N. Kayal, N. Saxena. *PRIMES is in P*. Annals of Mathematics 160 (2): 781–793. 2004.

<sup>2</sup>See <http://www.rsasecurity.com/rsalabs/node.asp?id=2093>.

## PRIME NUMBERS

Every natural number  $n \geq 2$  has at least two positive divisors: 1 and itself. If there are no other divisors, it is called a *prime* (number). Otherwise (i.e., if there exist other divisors), we talk about a *composite* (number).

In the subsequent paragraphs, we will usually denote primes by the letter  $p$ . The first few primes are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, ... (in particular, the number 1 is considered to be neither prime nor composite as it is a unit in the ring of the integers). As we will prove shortly, there are infinitely many primes. However, we have rather limited computational resources when it comes to determining whether a given number is prime: The number  $2^{82\,589\,933} - 1$ , which is the greatest known prime as of 2018, has only 24 862 048 digits, so its decimal representation would fit into many a prehistoric data storage device. Printing it as a book would, however, (assuming 60 rows on a page and 80 digits in a row) take 5 180 pages.

Now, let us introduce a theorem which gives a necessary and sufficient condition for being prime and is thus a fundamental ingredient for the proof of the unique factorization theorem.

**11.2.1. Theorem** (Euclid’s on primes). *An integer  $p \geq 2$  is a prime if and only if the following holds: for every pair of integers  $a, b$ ,  $p \mid ab$  implies that either  $p \mid a$  or  $p \mid b$  (or both).*

**PROOF.** “ $\implies$ ” Suppose that  $p$  is a prime and  $p \mid ab$ , where  $a, b \in \mathbb{Z}$ . Since  $(p, a)$  is a positive divisor of  $p$ , we have either  $(p, a) = p$  or  $(p, a) = 1$ . In the former case, we get  $p \mid a$ ; in the latter,  $p \mid b$  by part (2) of the previous lemma.

“ $\impliedby$ ” If  $p$  is not a prime, it has a positive divisor distinct from both 1 and  $p$ . Let us denote it by  $a$ . However, then we have  $b = \frac{p}{a} \in \mathbb{N}$  and  $p = ab$ , hence  $1 < a < p$ ,  $1 < b < p$ . We have thus found integers  $a, b$  such that  $p \mid ab$ , while  $p$  divides neither  $a$  nor  $b$ .  $\square$

**11.2.2. Fundamental theorem of arithmetic.** In arithmetic operations, we much rely on expressing the numbers (or more complicated objects) as products of simpler ones. The prime property distinguishes the simplest ones and we call them *irreducible*. We shall see in 12.3.5 how to define divisibility in an arbitrary integral domain (i.e., rings without divisors of zero). In some integral domains (e.g.  $\mathbb{Q}$ ), there are no elements with the prime property.



Other integral domains have such elements, yet they do not satisfy the unique factorization theorem. It is quite similar with the generalization of aforementioned Euclid’s theorem on primes – the elements which satisfy  $p \mid ab \implies (p \mid a \text{ or } p \mid b)$  are always irreducible, but the contrary is not generally true. Let us mention at least one example of an ambiguous

**11.B.5. Number of divisors and sum of divisors.** Prove the following formulae: Every integer  $a = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$  has got exactly

$$\tau(a) = (\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_k + 1)$$

positive divisors, which sum up to

$$\sigma(a) = \frac{p_1^{\alpha_1+1} - 1}{p_1 - 1} \cdots \frac{p_k^{\alpha_k+1} - 1}{p_k - 1}.$$

Moreover, let  $p_1, \dots, p_k$  be pairwise distinct primes and  $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k$  be non-negative integers. Denoting  $\gamma_i = \min\{\alpha_i, \beta_i\}$ ,  $\delta_i = \max\{\alpha_i, \beta_i\}$  for every  $i = 1, 2, \dots, k$ , then we have

$$(p_1^{\alpha_1} \cdots p_k^{\alpha_k}, p_1^{\beta_1} \cdots p_k^{\beta_k}) = p_1^{\gamma_1} \cdots p_k^{\gamma_k},$$

$$[p_1^{\alpha_1} \cdots p_k^{\alpha_k}, p_1^{\beta_1} \cdots p_k^{\beta_k}] = p_1^{\delta_1} \cdots p_k^{\delta_k}.$$

**Solution.** Every positive divisor of the integer  $a = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$  is of the form  $p_1^{\beta_1} \cdots p_k^{\beta_k}$ , where  $\beta_1, \dots, \beta_k \in \mathbb{N}_0$  and  $\beta_1 \leq \alpha_1, \beta_2 \leq \alpha_2, \dots, \beta_k \leq \alpha_k$ .

The justification of all of the claims is now a simple consequence of this explicit factorization description of the the divisors. To find the number of positive divisors, we employ elementary combinatorics (the product rule) to get  $\tau(a) = (\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_k + 1)$ .

Next, we can see that the formula for the sum of the divisors holds if we rewrite it in the form

$$(1 + p_1 + \cdots + p_1^{\alpha_1}) \cdots (1 + p_k + \cdots + p_k^{\alpha_k}),$$

realizing that every pair of parentheses contains the sum of a finite geometric series. The other statements follow directly from the definition.  $\square$

*The sum of all positive divisors of an integer is connected to the so-called perfect numbers. We say that an integer  $a$  is perfect if and only if  $\sigma(a) = 2a$ , i.e., if and only if the sum of all positive divisors of  $a$ , excluding  $a$  itself, equals  $a$ .*

*Perfect numbers include, for instance,  $6 = 1 + 2 + 3$ ,  $28 = 1 + 2 + 4 + 7 + 14$ ,  $496$ , and  $8128$  (this exhausts all perfect numbers less than 10 000).*

*It can be shown that even perfect numbers are in a tight relation with the so-called Mersenne primes since the following holds:*

**11.B.6.** Show that a natural number  $a$  is an even perfect number if and only if it is of the form  $a = 2^{q-1}(2^q - 1)$ , where  $2^q - 1$  is a prime.

**Solution.** If  $a = 2^{q-1}(2^q - 1)$ , where  $p = 2^q - 1$  is a prime, then the previous statement yields

$$\sigma(a) = \frac{2^q - 1}{2 - 1} \cdot (p + 1) = (2^q - 1) \cdot 2^q = 2a.$$

Such an integer  $a$  is thus a perfect number.

For the opposite direction, consider any even perfect number  $a$ , and let us write

$$a = 2^k \cdot m, \text{ where } m, k \in \mathbb{N} \text{ and } 2 \nmid m.$$

factorization – in  $\mathbb{Z}[\sqrt{-5}]$ , we have:<sup>3</sup>

$$6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5}).$$

However, it needs a longer discussion to verify that all of the mentioned factors are really irreducible in  $\mathbb{Z}[\sqrt{-5}]$ .

FUNDAMENTAL THEOREM OF ARITHMETIC

**Theorem.** Every natural number  $n$  can be expressed as the product of primes, and this expression is unique up to the order of the factors. (If  $n$  is a prime, then it is the “product” of a single prime  $n$ ; if  $n = 1$ , it is the empty product (the product of the empty set of primes)<sup>4</sup>)

**PROOF.** First, we prove by complete induction on  $n$  that every natural number  $n$  can be expressed as a product of primes. We have already discussed the validity of this statement for  $n = 1$ .

Now, let us suppose that  $n \geq 2$  and we have already proved that all natural numbers less than  $n$  can be factored to primes. If  $n$  is a prime, the statement is clearly true. If  $n$  is not a prime, then it has a divisor  $d$ ,  $1 < d < n$ . Denoting  $e = n/d$ , we also have  $1 < e < n$ . From the induction hypothesis, we get that both  $d$  and  $e$  can be expressed as products of primes, so their product  $d \cdot e = n$  can also be expressed in this way.

To prove uniqueness, let us have an equality of products  $n = p_1 \cdot p_2 \cdots p_s = q_1 \cdot q_2 \cdots q_t$ , where  $p_i, q_j$  are primes for all  $i \in \{1, \dots, s\}, j \in \{1, \dots, t\}$ . We will prove by induction on  $s$  that  $s = t$  and  $p_1 = q_1, \dots, p_s = q_s$ .

If  $s = 1$ , then  $p_1 = q_1 \cdots q_t$  is a prime. If we had  $t > 1$ , the integer  $p_1$  would have a divisor  $q_1$  such that  $1 < q_1 < p_1$  (since  $q_2 q_3 \cdots q_t > 1$ ), which is impossible. Therefore, we must have  $t = 1$  and  $p_1 = q_1$ .

Now, let us suppose that  $s \geq 2$  and the proposition holds for  $s - 1$ . It follows from the equality  $p_1 \cdot p_2 \cdots p_s = q_1 \cdot q_2 \cdots q_t$  that  $p_s$  divides the product  $q_1 \cdots q_t$ , which is, by Euclid’s theorem, possible only if  $p_s$  divides some  $q_j$  (and after relabeling we may assume  $j = t$ ). Since  $q_j$  is a prime, it follows that  $p_s = q_t$ . Dividing both sides of the original equality by this integer, we obtain  $p_1 \cdot p_2 \cdots p_{s-1} = q_1 \cdot q_2 \cdots q_{t-1}$ , and from the induction hypothesis, we get  $s - 1 = t - 1$ ,  $p_1 = q_1, \dots, p_{s-1} = q_{s-1}$ . Altogether, we have  $s = t$  and  $p_1 = q_1, \dots, p_{s-1} = q_{s-1}, p_s = q_s$ . This proves the uniqueness, and thus the entire theorem as well.  $\square$

**11.2.3. Prime distribution.** Don Zagier wrote: *There are two facts about the distribution of prime numbers. The first is that, [they are] the most arbitrary and ornery objects studied by mathematicians: they grow like weeds among the natural numbers, seeming to obey no other law than that of chance, and nobody can predict where the next one will sprout. The second fact*



<sup>3</sup>The symbol  $\mathbb{Z}[\sqrt{-5}]$  denotes the integers extended by a root of the equation  $x^2 = -5$ , which is defined similarly as we obtained the complex numbers by adjoining the number  $\sqrt{-1}$  to the reals.

Since the function  $\sigma$  is multiplicative (see 11.3.2), we have  $\sigma(a) = \sigma(2^k) \cdot \sigma(m) = (2^{k+1} - 1) \cdot \sigma(m)$ . However, it follows from  $a$  being perfect that  $\sigma(a) = 2a = 2^{k+1} \cdot m$ , whence

$$2^{k+1} \cdot m = (2^{k+1} - 1) \cdot \sigma(m).$$

Since  $2^{k+1} - 1$  is odd, we must have  $2^{k+1} - 1 \mid m$ , so we can lay  $m = (2^{k+1} - 1) \cdot n$  for an appropriate  $n \in \mathbb{N}$ . Rearranging leads to  $2^{k+1} \cdot n = \sigma(m)$ . Both  $m$  and  $n$  divide  $m$  (and since  $\frac{m}{n} = 2^{k+1} - 1 > 1$ , these integers are different), hence

$$2^{k+1} \cdot n = \sigma(m) \geq m + n = 2^{k+1} \cdot n,$$

and so  $\sigma(m) = m + n$ . However, this means that  $m$  is a prime with the sole divisors  $m$  and  $n = 1$ , whence  $a = 2^k \cdot (2^{k+1} - 1)$ , where  $2^{k+1} - 1 = m$  is a prime.  $\square$

**Remark.** On the other hand, people have been unsuccessful in describing odd perfect numbers; we even do not know whether there exists an odd perfect number.

Mersenne primes are those of the form  $2^k - 1$ . It should not go unnoticed that Mersenne primes are easily recognizable among all primes – for Mersenne numbers (excluding the primality requirement), there is a simple and fast procedure how to verify that they are primes. It is thus not by chance that the largest known primes are usually of the form  $2^k - 1$ .

Later, we will show how to efficiently verify whether a given Mersenne number is prime (see the Lucas-Lehmer test in part 11.6.9).

Although it may seem a weird and practically useless business to look for primes as great as possible, it pushes the borders of our cognition of mathematics forward and refines the used methods (as well as hardware). Moreover, the discoverers often benefit from this (Electronic Frontier Foundation issued EFF Cooperative Computing Awards for finding a prime having at least  $10^6$ ,  $10^7$ ,  $10^8$ , and  $10^9$  digits – rewards of 50 and 100 dollars, respectively, for the first and second categories were paid in 2000 and 2009, respectively, to the GIMPS project in both cases. Apparently, it will take a while before the other prizes are awarded.)

### C. Congruences



In this paragraph, we will see in practice how wielding basic operations with congruences can improve the expressing of our reasonings about various problems: We would be able to solve them without congruences, using only the basic properties of divisibility. However, with the help of congruences, our considerations will often be much shorter and clearer.

**11.C.1.** Show that, For any  $a, b \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ , the following conditions are equivalent:

- i)  $a \equiv b \pmod{m}$ ,
- ii)  $a = b + mt$  for an appropriate  $t \in \mathbb{Z}$ ,
- iii)  $m \mid a - b$ .

is even more astonishing, for it states just the opposite: that the prime numbers exhibit stunning regularity, that there are laws governing their behavior, and that they obey these laws with almost military precision.

In the next paragraphs, we will discuss the following questions:

*Are there infinitely many primes?*

*Are there infinitely many primes in every (or at least one) arithmetic sequence?*

*How are the primes distributed among the natural numbers?*

The first question has got an easy answer and the fundamental theorem was known to Euclid already around 300 BC:

**Theorem (Euclid).** *There are infinitely many primes among the natural numbers.*

**PROOF.** Suppose that there are only finitely many, and let them be denoted by  $p_1, p_2, \dots, p_n$ . Set  $N = p_1 \cdot p_2 \cdot \dots \cdot p_n + 1$ . This integer, being greater than 1, is either a prime or it is divisible by a prime different from  $p_1, \dots, p_n$  (since the primes  $p_1, \dots, p_n$  divide the integer  $N - 1$ ), which is a contradiction.  $\square$

**11.2.4.** Next, we will mention a rather strong statement, whose proof is very laborious (that is why we do not present it), yet it can be done by elementary means<sup>5</sup>.

**Theorem (Chebyshev’s, Bertrand’s postulate).** *For every integer  $n > 1$ , there is at least one prime  $p$  satisfying  $n < p < 2n$ .*

**11.2.5.** Primes are distributed quite uniformly in the sense that in any “reasonable” arithmetic sequence (i.e. such that its terms are coprime), there are infinitely many of them.

For instance, considering the remainders upon division by 4, there are infinitely many primes with remainder 1 as well as infinitely many primes with remainder 3 (of course, there is no prime with remainder 0 and only one prime with remainder 2). The situation is analogous for remainders upon division by other integers, as explained by the following theorem, whose proof is very difficult.

**Theorem (Dirichlet’s on primes).** *If  $a, m$  are coprime natural numbers, there are infinitely many primes  $k$  such that  $mk + a$  is a prime. In other words, there are infinitely many primes among the integers  $1 \cdot m + a, 2 \cdot m + a, 3 \cdot m + a, \dots$*

We can at least mention a proof of a special case of this theorem, which is a modification of Euler’s proof of the infinitude of primes.

**Proposition.** *There are infinitely many primes of the form  $4k + 3$ , where  $k \in \mathbb{N}_0$ .*

<sup>5</sup>See Wikipedia, *Proof of Bertrand’s postulate*, [http://en.wikipedia.org/wiki/Proof\\_of\\_Bertrand’s\\_postulate](http://en.wikipedia.org/wiki/Proof_of_Bertrand’s_postulate) (as of July 29, 2017) or see M. Aigner, G. Ziegler, *Proofs from THE BOOK*, Springer, 2009.

**Solution.** (1)  $\implies$  (3) If  $a = q_1m + r$ ,  $b = q_2m + r$ , then  $a - b = (q_1 - q_2)m$ .

(3)  $\implies$  (2) If  $m \mid a - b$ , then there is a  $t \in \mathbb{Z}$  such that  $m \cdot t = a - b$ , i.e.,  $a = b + mt$ .

(2)  $\implies$  (1) If  $a = b + mt$ , then, expressing  $b = mq + r$ , it follows that  $a = m(q + t) + r$ . Therefore,  $a$  and  $b$  share the same remainder  $r$  upon division by  $m$ , i.e.,  $a \equiv b \pmod{m}$ .  $\square$

**11.C.2.** Prove fundamental properties of congruences stated in 11.3.1.

**Solution.**

i) If  $a \equiv b \pmod{m}$  and  $c \equiv d \pmod{m}$ , by the previous lemma, there are integers  $s, t$  such that  $a = b + ms$ ,  $c = d + mt$ . However, then we have  $a + c = b + d + m(s + t)$ , and, by the lemma again,  $a + c \equiv b + d \pmod{m}$ .

Adding a congruence  $a \equiv b \pmod{m}$  to  $mk \equiv 0 \pmod{m}$ , which is clearly valid, leads to  $a + mk \equiv b \pmod{m}$ .

ii) If  $a \equiv b \pmod{m}$  and  $c \equiv d \pmod{m}$ , there are integers  $s, t$  such that  $a = b + ms$ ,  $c = d + mt$ . Then,

$$ac = (b + ms)(d + mt) = bd + m(bt + ds + mst),$$

whence we get  $ac \equiv bd \pmod{m}$ .

iii) Let  $a \equiv b \pmod{m}$  and  $n$  be a natural number. Since

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \dots + b^{n-1}),$$

it follows that  $a^n \equiv b^n \pmod{m}$  as well.

iv) Suppose that  $a \equiv b \pmod{m}$ ,  $a = a_1 \cdot d$ ,  $b = b_1 \cdot d$ , and  $(m, d) = 1$ . By the lemma, the difference  $a - b = (a_1 - b_1) \cdot d$  is divisible by the integer  $m$ , and since  $(m, d) = 1$ , the integer  $a_1 - b_1$  is also divisible by  $m$  (by lemma 11.1.5). Hence it follows that  $a_1 \equiv b_1 \pmod{m}$ . Further, if  $ad \equiv bd \pmod{md}$ , i.e.,  $md \mid ad - bd$ , we get directly from the definition of divisibility that  $m \mid a - b$ .

v) If  $a \equiv b \pmod{m}$ , then  $a - b$  is a multiple of  $m$ , and hence a multiple of any divisor  $d$  of  $m$ , so  $a \equiv b \pmod{d}$ .

vi) Suppose that  $a \equiv b \pmod{m}$ ,  $b = b_1d$ ,  $m = m_1d$ . Then there is an integer  $t$  such that  $a = b + mt = b_1d + m_1dt = (b_1 + m_1t)d$ , hence  $d \mid a$ .

vii) If  $a \equiv b \pmod{m_1}$ ,  $a \equiv b \pmod{m_2}, \dots, a \equiv b \pmod{m_k}$ , then the difference  $a - b$  is a common multiple of the integers  $m_1, m_2, \dots, m_k$ , and so it is divisible by their least common multiple  $[m_1, m_2, \dots, m_k]$  whence it follows that  $a \equiv b \pmod{[m_1, \dots, m_k]}$ .

**Remark.** We have already used some properties of congruences without explicitly mentioning it – now, the result of the exercise 11.A.3 can be reformulated as “if  $a \equiv 1 \pmod{m}$ ,  $b \equiv 1 \pmod{m}$ , then also  $ab \equiv 1 \pmod{m}$ ”, which is a special case of item (2) of the previous theorem.

It is not by chance because any statement which uses congruences can be reformulated in terms of divisibility. The usefulness of congruences thus lies not in the strength to solve more problems than without them, but rather in being a very

**PROOF.** Suppose the contrary, i.e., there are only finitely many primes of this form, and let them be denoted by  $p_1 = 3, p_2 = 7, p_3 = 11, \dots, p_n$ . Further, let us set  $N = 4p_2 \cdot p_3 \cdot \dots \cdot p_n + 3$ . Factoring  $N$ , the product must contain (according to the result of exercise 11.A.3) at least one prime  $p$  of the form  $4k + 3$ . If not,  $N$  would be a product of only primes of the form  $4k + 1$ , so  $N$  as well would give remainder 1 upon division by 4, which is not true. However, none of the primes  $p_1, p_2, \dots, p_n$  can play the role of the mentioned  $p$  since if we had  $p_i \mid N$  for some  $i \in \{2, \dots, n\}$ , then we would get  $p_i \mid 3$ . Similarly,  $3 \nmid N$ , and we thus get a contradiction with the assumption of finitely many primes of the given form.  $\square$

An analogous elementary proof can be used for primes of the form  $3k + 2$  or  $6k + 5$ ; however, it will not work for primes of the form  $3k + 1$  or  $4k + 1$  (think this out well!). We will be able to remedy this in the latter case in part 11.4.11 about quadratic congruences).

**11.2.6.** From the propositions mentioned in this chapter, one can roughly imagine how “dense” the primes appear among the natural numbers. It is more accurately described (although “only” asymptotically) by the following, very important theorem, which was proved independently by J. Hadamard and Ch. J. de la Vallée-Poussin in 1896 (but we shall not go into the proof here).

**Theorem (Prime Number Theorem).** Let  $\pi(x)$  denote the number of primes less than or equal to a number  $x \in \mathbb{R}$ . Then

$$\pi(x) \sim \frac{x}{\ln x},$$

i.e., the quotient of the functions  $\pi(x)$  and  $x / \ln x$  approaches 1 for  $x \rightarrow \infty$ .

The following table illustrates how good the asymptotic estimate  $\pi(x) \sim x / \ln(x)$  is in several concrete instances in reality:

$x$	$\pi(x)$	$x / \ln(x)$	relative error
100	25	21.71	0.13
1000	168	144.76	0.13
10000	1229	1085.73	0.11
100000	9592	8685.88	0.09
500000	41538	38102.89	0.08

The density of primes among the natural numbers is also partially described by the following result by Euler. Let us remind the formula  $\sum_{n \in \mathbb{N}} \frac{1}{n^2} = \frac{\pi^2}{6}$  (see 7.1.10) which indicates that in view of the next proposition the primes are distributed more “densely” in  $\mathbb{N}$  than squares.

**Proposition.** Let  $P$  denote the set of all primes, then

$$\sum_{p \in P} \frac{1}{p} = \infty.$$

convenient way of writing which simplifies both expressions and reasonings.

**11.C.3.**

- i) Find the remainder of the integer  $7^{30}$  when divided by 50.
- ii) Find the last two digits of the decimal representation of the integer  $7^{30}$ .

**Solution.**

- i) Since  $7^2 = 49 \equiv -1 \pmod{50}$ , using the properties of congruences, which are mentioned above,  $7^{30} \equiv (-1)^{15} = -1 \pmod{50}$ , so the remainder of  $7^{30}$  upon division by 50 is 49.
- ii) Our task is actually to determine the remainder of  $7^{30}$  upon division by 100. We know from the above that the integer  $7^{30}$  leaves remainder 49 when divided by 50, so the last two digits are either 49 or 99. In particular, we already know that  $7^{30} \equiv -1 \pmod{25}$ , and we can easily calculate that  $7^{30} \equiv (-1)^{30} = 1 \pmod{4}$ . Since  $(4, 25) = 1$ , the wanted pair of digits is 49 (it leaves the desired remainder upon division by both 25 and 4).  $\square$

We now show how helpful the notion of congruence can be in solving problems similar to one already solved in 11.A.8 (where the induction and binomial theorem were used).



**11.C.4.** Prove that for any  $n \in \mathbb{N}$ , the integer  $37^{n+2} + 16^{n+1} + 23^n$  is divisible by 7.

**Solution.** We have  $37 \equiv 16 \equiv 23 \equiv 2 \pmod{7}$ , so by the basic properties of congruences,

$$\begin{aligned} 37^{n+2} + 16^{n+1} + 23^n &\equiv 2^{n+2} + 2^{n+1} + 2^n \\ &= 2^n(4+2+1) \equiv 0 \pmod{7}. \quad \square \end{aligned}$$

**11.C.5.** Prove that the integer  $n = (835^5 + 6)^{18} - 1$  is divisible by 112.

**Solution.** We factor  $112 = 7 \cdot 16$ . Since  $(7, 16) = 1$ , it suffices to show that  $7 \mid n$  and  $16 \mid n$ . We have  $835 \equiv 2 \pmod{7}$ , so

$$\begin{aligned} n &\equiv (2^5 + 6)^{18} - 1 = 38^{18} - 1 \equiv 3^{18} - 1 \\ &= 27^6 - 1 \equiv (-1)^6 - 1 = 0 \pmod{7}, \end{aligned}$$

hence  $7 \mid n$ . Similarly,  $835 \equiv 3 \pmod{16}$ , so

$$\begin{aligned} n &\equiv (3^5 + 6)^{18} - 1 = (3 \cdot 81 + 6)^{18} - 1 \equiv (3 \cdot 1 + 6)^{18} - 1 \\ &= 9^{18} - 1 = 81^9 - 1 \equiv 1^9 - 1 = 0 \pmod{16}, \end{aligned}$$

hence  $16 \mid n$ . Altogether,  $112 \mid n$ , which was to be proved.  $\square$

**11.C.6.** Prove that the following relations hold for prime  $p$ :

- i) If  $k \in \{1, \dots, p-1\}$ , then  $p \mid \binom{p}{k}$ .
- ii) If  $a, b \in \mathbb{Z}$ , then  $a^p + b^p \equiv (a+b)^p \pmod{p}$ .

**Solution.**

**PROOF.** Let  $n$  be an arbitrary natural number and  $p_1, \dots, p_{\pi(n)}$  all primes less than or equal to  $n$ . Let us set

$$\lambda(n) = \prod_{i=1}^{\pi(n)} \left(1 - \frac{1}{p_i}\right)^{-1}.$$

The particular factors can be perceived as sums of geometric series, hence

$$\lambda(n) = \prod_{i=1}^{\pi(n)} \left(\sum_{\alpha_i=0}^{\infty} \frac{1}{p_i^{\alpha_i}}\right) = \sum \frac{1}{p_1^{\alpha_1} \cdots p_{\pi(n)}^{\alpha_{\pi(n)}}},$$

where we sum over all  $\pi(n)$ -tuples of non-negative integers  $(\alpha_1, \dots, \alpha_{\pi(n)})$ . Since every integer not exceeding  $n$  factors to only primes in the set  $\{p_1, \dots, p_{\pi(n)}\}$ , all of them are included in this sum. Therefore,  $\lambda(n) > 1 + \frac{1}{2} + \dots + \frac{1}{n}$ , and since the harmonic series is divergent (see 5.D.1), we also have  $\lim_{n \rightarrow \infty} \lambda(n) = \infty$ .

Taking into account the expansion of the function  $\ln(1+x)$  to a power series (see 6.D.47), we further get

$$\begin{aligned} \ln \lambda(n) &= - \sum_{i=1}^{\pi(n)} \ln \left(1 - \frac{1}{p_i}\right) = \sum_{i=1}^{\pi(n)} \sum_{m=1}^{\infty} (mp_i^m)^{-1} = \\ &= p_1^{-1} + \dots + p_{\pi(n)}^{-1} + \sum_{i=1}^{\pi(n)} \sum_{m=2}^{\infty} (mp_i^m)^{-1}. \end{aligned}$$

Since the inner sum can be bound from above by

$$\begin{aligned} \sum_{m=2}^{\infty} (mp_i^m)^{-1} &< \sum_{m=2}^{\infty} p_i^{-m} = \\ &= p_i^{-2} (1 - p_i^{-1})^{-1} \leq 2p_i^{-2}, \end{aligned}$$

the divergent sequence  $\ln \lambda(n) < \sum_{i=1}^{\pi(n)} p_i^{-1} + 2 \sum_{i=1}^{\pi(n)} p_i^{-2}$  can also be bounded from above. The second sum apparently converges (since the series  $\sum_{n=1}^{\infty} n^{-2}$  does), so the first sum  $\sum_{i=1}^{\pi(n)} p_i^{-1}$  must diverge, which is what we wanted to prove.  $\square$

**3. Congruences and basic theorems**

This concept was introduced by C. F. Gauss in 1801 in his book *Disquisitiones Arithmeticae*. Although being a simple one, its contribution to number theory is mainly in making some reasonings (even quite complicated ones) be written much more compactly and transparently.



CONGRUENCE

If integers  $a, b$  give the same remainder  $r$  (where  $0 \leq r < m$ ) when divided by a natural number  $m$ , we say that they are *congruent modulo  $m$*  and write it as

$$a \equiv b \pmod{m}.$$

In the other case, we say that the integers  $a, b$  are not congruent modulo  $m$ , writing

$$a \not\equiv b \pmod{m}.$$



i) Since the binomial coefficient satisfies

$$\binom{p}{k} = \frac{p(p-1)\cdots(p-k+1)}{k!},$$

which is an integer, we hence know that  $k!$  divides the product  $p(p-1)\cdots(p-k+1)$ . However, since the integer  $k!$  is coprime to the prime  $p$ , we thus get that  $k! \mid (p-1)\cdots(p-k+1)$ , whence it follows that  $p \mid \binom{p}{k}$ .

ii) The binomial theorem implies that

$$(a+b)^p = a^p + \binom{p}{1}a^{p-1}b + \cdots + \binom{p}{p-1}ab^{p-1} + b^p.$$

Thanks to the previous item, we have  $\binom{p}{k} \equiv 0 \pmod{p}$  for any  $k \in \{1, \dots, p-1\}$ , whence the statement follows easily.  $\square$

**11.C.7.** Prove that for any natural numbers  $m, n$  and any integers  $a, b$  such that  $a \equiv b \pmod{m^n}$ , it is true that

$$a^m \equiv b^m \pmod{m^{n+1}}.$$

**Solution.** Since clearly  $m \mid m^n$ , we get that the congruence  $a \equiv b \pmod{m}$  holds, invoking property (5) of 11.3.1. Therefore, considering the algebraic identity

$$a^m - b^m = (a-b)(a^{m-1} + a^{m-2}b + \cdots + ab^{m-2} + b^{m-1}),$$

all the summands in the second pair of parentheses are congruent to  $a^{m-1}$  modulo  $m$ , so

$$a^{m-1} + a^{m-2}b + \cdots + b^{m-1} \equiv m \cdot a^{m-1} \equiv 0 \pmod{m}.$$

Since  $m^n$  divides  $a-b$ , and the sum  $a^{m-1} + a^{m-2}b + \cdots + b^{m-1}$  is divisible by  $m$ , we get that  $m^{n+1}$  must divide their product, which means that  $a^m \equiv b^m \pmod{m^{n+1}}$ .  $\square$

**11.C.8.** Using the result of the previous exercise (see also 11.A.2), prove that:

- i) integers  $a$  which are not divisible by 3 satisfy  $a^3 \equiv \pm 1 \pmod{9}$ ,
- ii) odd integers  $a$  satisfy  $a^4 \equiv 1 \pmod{16}$ .

**Solution.**

- i) Cubing the congruence  $a \equiv \pm 1 \pmod{3}$  (and, again, raising the exponent of the modulus), we get  $a^3 \equiv \pm 1 \pmod{3^2}$ .
- ii) This statement was proved already in the third part of exercise 11.A.2. Now, we will present another proof. Thanks to part (ii) of the mentioned exercise, we know that every odd integer  $a$  satisfies  $a^2 \equiv 1 \pmod{2^3}$ . Squaring this (and recalling the above exercise) leads to  $a^4 \equiv 1^2 \pmod{2^4}$ .

**11.C.9. Divisibility rules.** We can surely recall the basic rules of divisibility (at least by the numbers 2, 3, 4, 5, 6, 9 or 10) in terms of the decimal representation of a given integer. However, how can these rules be proved and can they be extended to other divisors as well?

We already know that we can restrict ourselves to divisibility by powers of primes (for instance, divisibility by 6 can be tested using divisibility by 2 and 3).



Whenever it is apparent that we are working with congruence relations, we usually omit the symbol  $\pmod{m}$ , or write just  $a \equiv b \pmod{m}$ .

**11.3.1. Fundamental properties.** It follows directly from the definition that the congruence modulo  $m$  is an *equivalence* relation.

Now, we will state further properties of congruences which are proven in detail in the practical column, see 11.C.2.

PROPERTIES OF CONGRUENCES

- (1) Congruences with respect to the same modulus can be added. An arbitrary multiple of the modulus can be added to either side.
- (2) Congruences with respect to the same modulus can be multiplied.
- (3) Both sides of a congruence can be raised to the power of the same natural number.
- (4) Both sides of a congruence can be divided by their common divisor provided it is coprime to the modulus. Both sides of a congruence together with the modulus can be divided by a positive divisor common to all three of them.
- (5) If a congruence is valid with respect to a modulus  $m$ , it is also valid with respect to any modulus  $d$  which divides  $m$ .
- (6) If either side of a congruence and the modulus are divisible by an integer, then this integer must divide the other side of the congruence as well.
- (7) If a congruence is valid with respect to moduli  $m_1, \dots, m_k$ , it is also valid with respect to their least common multiple  $[m_1, \dots, m_k]$ .

**Remark.** We have already used some properties of congruences without explicitly mentioning it – now, the result of the exercise 11.A.3 can be reformulated as “if  $a \equiv 1 \pmod{m}$ ,  $b \equiv 1 \pmod{m}$ , then also  $ab \equiv 1 \pmod{m}$ ”, which is a special case of the property (2) above.

It is not by chance because any statement which uses congruences can be reformulated in terms of divisibility. The usefulness of congruences thus lies not in the strength to solve more problems than without them, but rather in being a very convenient way of writing, which simplifies both expressions and reasonings.

**11.3.2. Arithmetic functions.** By arithmetic function we mean any function whose domain is the set of natural numbers. Good examples are the number of divisors  $\tau(a)$  or their sum  $\sigma(a)$ , computed in 11.B.5. A prominent example is the following function counting the coprime numbers less than given  $a$ :



The rule for divisibility by 9 says that a given integer is divisible by 9 if and only if its digit sum is. We will prove this as a consequence of a much stronger statement: It holds that every integer is congruent to its digit sum modulo 9 (in particular, it is congruent to zero if and only if its digit sum is). And this is trivial to prove: The digit sum of an integer  $n = a_k 10^k + a_{k-1} 10^{k-1} + \dots + a_1 10 + a_0$  is equal to  $S(n) = a_k + a_{k-1} + \dots + a_0$ , and since  $10^\ell \equiv 1^\ell = 1 \pmod{9}$  for any  $\ell \in \mathbb{N}_0$ , we get

$$n = a_k 10^k + \dots + a_0 \equiv a_k + \dots + a_0 = S(n) \pmod{9}.$$

This derivation is valid also if we replace 9 with 3.

The rule for divisibility by 11, which we have not mentioned yet, works similarly. Here, we have  $10^\ell \equiv (-1)^\ell \pmod{11}$ , so we get

$$\begin{aligned} n &= a_k 10^k + \dots + a_0 \equiv a_k (-1)^k + \dots + a_1 (-1) + a_0 \\ &\equiv (a_0 + a_2 + \dots) - (a_1 + a_3 + \dots) \pmod{11}. \end{aligned}$$

Therefore, an integer is divisible by 11 if and only if the difference of the sum of the digits at even places and the sum of the digits at odd places is.

There is a nice trick for the divisors 7 and 13: We have  $1001 = 7 \cdot 11 \cdot 13$ ; an integer  $n = 1000a + b$  thus satisfies  $n \equiv -a + b \pmod{m}$ , where  $m$  is any of the numbers 7, 11, 13. Therefore, 2015 is divisible by 13 since  $015 - 2 = 13$ . Similarly, 2016 is divisible by 7 as  $016 - 2 = 14$  is a multiple of 7. We could justify that 2013 is a multiple of 11 in the same way, but the aforementioned criterion  $11 \mid (3 + 0) - (1 + 2)$  is maybe more smart.

**Using divisibility for error detection.** Let us note that divisibility by eleven is often used for making decimal codes which allow us to detect a single-digit error. If we make such a mistake when copying an integer which is divisible by eleven, then the resulting integer is surely not (see the aforementioned criterion of divisibility by eleven). More details can be found in chapter 12.5.1 about coding. For instance, the national identification numbers in the Czech Republic and Slovakia contain a check digit which completes the code into an integer divisible by eleven.

Similarly, the numbers of bank accounts managed by Czech banks must comply with a similar (only a bit more complicated) procedure. Both the transformed 6-digit prefix  $a_5 a_4 a_3 a_2 a_1 a_0$  and the 10-digit account number  $b_9 b_8 b_7 b_6 b_5 b_4 b_3 b_2 b_1 b_0$  must satisfy the following condition on divisibility by eleven (here, we mention only the one for the number without the prefix):

$$\begin{aligned} 0 &\equiv b_9 2^9 + b_8 2^8 + b_7 2^7 + \dots + b_3 2^3 + b_2 2^2 + b_1 2^1 + b_0 2^0 \\ &\equiv -5b_9 + 3b_8 - 4b_7 - 2b_6 - b_5 \\ &\quad + 5b_4 - 3b_3 + 4b_2 + 2b_1 + b_0 \pmod{11}. \end{aligned}$$

This condition can be shortly described so that the account number, perceived as being in binary (though with usage of decimal digits) is to be divisible by eleven.

EULER'S TOTIENT FUNCTION  $\varphi$

For a natural number, we define the value of Euler's totient function  $\varphi$  as

$$\varphi(n) = |\{a \in \mathbb{N} \mid 0 < a \leq n, (a, n) = 1\}|.$$

For example,  $\varphi(1) = 1, \varphi(5) = 4, \varphi(6) = 2$ . If  $p$  is a prime, then clearly  $\varphi(p) = p - 1$  (all natural numbers less than  $p$  are coprime to it).

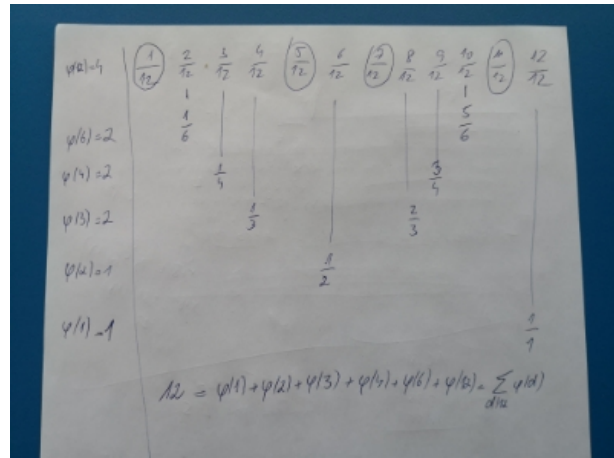
We are going to prove several important properties of the function  $\varphi$ . Let us start with a quite simple observation:

**Lemma.** Let  $n \in \mathbb{N}$ . Then,  $\sum_{d|n} \varphi(d) = n$ .

PROOF. Let us consider the  $n$  fractions

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}.$$

Reducing them to lowest terms and grouping them together by the denominators, we get just the statement in question (see the picture illustrating the case  $n = 12$ ).  $\square$



**11.3.3.** Soon we shall recognize, that knowing the values of the Euler's totient function may be extremely useful. The following explicit formula is crucial. We might prove it by the inclusion-exclusion principle, determining the number of integers not coprime to  $n$  in a given interval. Here we shall employ a more conceptual approach which is of great interest by itself, and thus, the proof is postponed to 11.3.8 below.

**Theorem.** Let  $n \in \mathbb{N}$  factor to primes as  $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$ . Then,

$$\varphi(n) = n \cdot \left(1 - \frac{1}{p_1}\right) \dots \left(1 - \frac{1}{p_k}\right).$$

**11.3.4. Multiplicative functions.** Dealing with arithmetic functions, we introduce the concept of "multiplicativity" which concerns only the coprime arguments:

**Definition.** A multiplicative function on the natural numbers is such an arithmetic function which, for all pairs of coprime natural numbers  $a, b$ , satisfies

$$f(a \cdot b) = f(a) \cdot f(b).$$

**11.C.10.** Verify that the account number of the Masaryk university, 85636621, is built correctly.

**Solution.** We will test the condition of divisibility by eleven:  
 $-5b_9 + 3b_8 - 4b_7 - 2b_6 - b_5 + 5b_4 - 3b_3 + 4b_2 + 2b_1 + b_0$   
 $\equiv -4 \cdot 8 - 2 \cdot 5 - 1 \cdot 6 + 5 \cdot 3 - 3 \cdot 6 + 4 \cdot 6 + 2 \cdot 2 + 1 \cdot 1$   
 $\equiv 0 \pmod{11}.$   $\square$

**Euler's totient function.** The totient function  $\varphi$  assigns to a natural number  $m$  the number of natural numbers which are less than or equal to  $m$  and coprime to  $m$ , which can be written as  $\varphi(m) = |\{a \in \mathbb{N} \mid 0 < a \leq m, (a, m) = 1\}|$ .

However, to be able to evaluate it efficiently, one needs to know the factorization of the input integer  $m$  to primes. In such a case, for  $m = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ , we have

$$\varphi(m) = (p_1 - 1)p_1^{\alpha_1 - 1} \cdots (p_k - 1)p_k^{\alpha_k - 1}.$$

In particular, we know that  $\varphi(p^\alpha) = (p - 1) \cdot p^{\alpha - 1}$  and that  $\varphi(m \cdot n) = \varphi(m) \cdot \varphi(n)$  holds whenever  $m, n$  are coprime.

**11.C.11.** Calculate  $\varphi(72)$ .

**Solution.**  $72 = 2^3 \cdot 3^2 \implies \varphi(72) = 72 \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{3}) = 24$ , alternatively  $\varphi(72) = \varphi(8) \cdot \varphi(9) = 4 \cdot 6 = 24$ .  $\square$

**11.C.12.**

- i) Determine all natural numbers  $n$  for which  $\varphi(n)$  is odd.
- ii) Prove that  $\forall n \in \mathbb{N} : \varphi(4n + 2) = \varphi(2n + 1)$ .

**Solution.**

- i) We clearly have  $\varphi(1) = \varphi(2) = 1$ . Every integer  $n \geq 3$  is either divisible by an odd prime  $p$  (then,  $\varphi(n)$  is divisible  $p - 1$ , which is an even integer) or  $n$  is a (higher-than-first) power of two (and then,  $\varphi(2^\alpha) = 2^{\alpha - 1}$  is even as well). Altogether, we have found out that  $\varphi(n)$  is odd only for  $n = 1, 2$ .
- ii) The integer  $2n + 1$  is odd, so  $(2, 2n + 1) = 1$ , and hence  $\varphi(4n + 2) = \varphi(2 \cdot (2n + 1)) = \varphi(2) \cdot \varphi(2n + 1) = \varphi(2n + 1)$ .  $\square$

**11.C.13.** Find all natural numbers  $m$  for which:



- i)  $\varphi(m) = 30$ ,
- ii)  $\varphi(m) = 34$ ,
- iii)  $\varphi(m) = 20$ ,
- iv)  $\varphi(m) = \frac{m}{3}$ .

**Solution.** In all of the above cases, we are looking for the fibers of a given integer  $a$  in the form  $m = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ , and we proceed as follows:

- Since  $\varphi(m) = (p_1 - 1)p_1^{\alpha_1 - 1} \cdots (p_k - 1)p_k^{\alpha_k - 1} = a$ , every prime  $p$  which divides  $m$  must satisfy

$$p - 1 \mid a.$$

- Similarly, every prime  $p$  whose higher power divides  $m$  must divide  $a$ . More exactly, we must even have  $p^{\alpha - 1} \mid a$ .

Clearly, the multiplicative functions include, for instance, the number of divisors  $\tau(n)$ , or their sum  $\sigma(n)$ , cf. 11.B.5.

It follows directly from Theorem 11.3.3 that the Euler's totient function is a multiplicative arithmetic function, too:

**Corollary.** Let  $a, b \in \mathbb{N}$ ,  $(a, b) = 1$ . Then

$$\varphi(a \cdot b) = \varphi(a) \cdot \varphi(b).$$

**Remark.** The multiplicativity of  $\varphi$  can also be derived directly from the knowledge that

$$(n, ab) = 1 \iff (n, a) = 1 \wedge (n, b) = 1.$$

Then the easy fact about the totient function on powers of primes

$$\varphi(p^\alpha) = p^\alpha - p^{\alpha - 1} = (p - 1) \cdot p^{\alpha - 1},$$

leads to the formula for the computation of  $\varphi$  in yet another alternative way.

**11.3.5. Möbius function.** Let us start our detour leading to the proof of Theorem 11.3.3. At the same time, the introduced convolution and inversion formulae should provide a glance into a very important conceptual direction.



**Definition.** Let a natural number  $n$  be factored to distinct primes:  $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ . The value of the *Möbius function*  $\mu(n)$  is defined to be 0 if  $\alpha_i > 1$  for some  $i$  (i.e., if  $n$  is divisible by a square), and  $(-1)^k$  otherwise. Further, we define  $\mu(1) = 1$  (in accordance with the convention that 1 factors to the product of zero primes).

For instance,  $\mu(4) = \mu(2^2) = 0$ ,  $\mu(6) = \mu(2 \cdot 3) = (-1)^2 = 1$ ,  $\mu(2) = \mu(3) = -1$ . By the very definition,  $\mu$  is a multiplicative function, too.

In the next paragraphs, we prove several important properties of the Möbius function, especially the so-called *Möbius inversion formula*. Let us start with a simple observation:

**Lemma.** For all  $n \in \mathbb{N} \setminus \{1\}$ , it holds that  $\sum_{d \mid n} \mu(d) = 0$ .

**PROOF.** Writing  $n$  as  $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ , then all divisors  $d$  of  $n$  are of the form  $d = p_1^{\beta_1} \cdots p_k^{\beta_k}$ , where  $0 \leq \beta_i \leq \alpha_i$  for all  $i \in \{1, \dots, k\}$ . Therefore,

$$\begin{aligned} \sum_{d \mid n} \mu(d) &= \sum_{\substack{(\beta_1, \dots, \beta_k) \in (\mathbb{N} \cup \{0\})^k \\ 0 \leq \beta_i \leq \alpha_i}} \mu(p_1^{\beta_1} \cdots p_k^{\beta_k}) \\ &= \sum_{(\beta_1, \dots, \beta_k) \in \{0, 1\}^k} \mu(p_1^{\beta_1} \cdots p_k^{\beta_k}) \\ &= \binom{k}{0} + \binom{k}{1} \cdot (-1) + \binom{k}{2} \cdot (-1)^2 + \cdots + \binom{k}{k} \cdot (-1)^k \\ &= (1 + (-1))^k = 0. \end{aligned}$$

In the third equality, we used a combinatorial reasoning – the summand  $\binom{k}{\ell} (-1)^\ell$  gives the contribution of the divisors  $d = p_1^{\beta_1} \cdots p_k^{\beta_k}$  with the property that exactly  $\ell$  of the exponents  $\beta_1, \dots, \beta_k$  are equal to one; there are  $\binom{k}{\ell}$  of them, and each satisfies that  $\mu(p_1^{\beta_1} \cdots p_k^{\beta_k}) = (-1)^\ell$ .  $\square$

- This procedure results in a finite set of candidates for  $m$ , which can be eliminated in a convenient way, sometimes also using the fact that any prime dividing  $a$  must occur in the factorization of  $\varphi(m)$  as a divisor of some  $p - 1$  or in a prime power  $p^{\alpha-1}$ .

Now, let us solve problems i)-iii):

- i) Every prime  $p$  from the factorization of  $m$  must satisfy  $p - 1 \mid 30$ , so  $p - 1 \in \{1, 2, 3, 5, 6, 10, 15, 30\}$ , which is satisfied by primes  $p \in \{2, 3, 7, 11, 31\}$ , and only 2 and 3 of them can divide  $m$  in higher power than the first. Therefore,

$$m = 2^\alpha 3^\beta 7^\gamma 11^\delta 31^\varepsilon,$$

where  $\alpha, \beta \in \{0, 1, 2\}, \gamma, \delta, \varepsilon \in \{0, 1\}$ . The analysis of the possibilities can be further simplified if we realize that  $\varphi(3) = 2, \varphi(3^2) = \varphi(7) = 6, \varphi(11) = 10$  are all integers which divide 30 into an odd integer greater than 1. Therefore, if we had, for instance,  $m = 7 \cdot m_1$ , where  $7 \nmid m_1$ , then we would also have  $\varphi(m_1) = 5$ , which is impossible, as we know from the previous exercise.

We thus get  $\beta = \gamma = \delta = 0$  and  $m = 2^\alpha \cdot 31^\varepsilon$ , whence we can easily obtain the solution  $m \in \{31, 62\}$ .

- ii) Similarly to above, only primes  $p \in \{2, 3\}$  can divide  $m$ , and the prime 3 can divide  $m$  only in the first power. However, since  $\frac{34}{\varphi(3)} = 17$ , the prime 3 cannot divide  $m$  at all. The remaining possibility,  $m = 2^\alpha$ , leads to  $34 = 2^{\alpha-1}$ , which is also impossible. Therefore, there is no such number  $m$ .
- iii) Now, every prime  $p$  dividing  $m$  must satisfy  $p - 1 \mid 20$ , so  $p - 1 \in \{1, 2, 4, 5, 10, 20\}$ , which is satisfied by primes  $p \in \{2, 3, 5, 11\}$ , and only 2 and 5 of those can divide  $m$  in higher power. We thus have

$$m = 2^\alpha 3^\beta 5^\gamma 11^\delta,$$

where  $\alpha \in \{0, 1, 2, 3\}, \gamma \in \{0, 1, 2\}, \beta, \delta \in \{0, 1\}$ .

First, consider  $\delta = 1$ . Then,  $\varphi(2^\alpha 3^\beta 5^\gamma) = 2$ , whence we easily get that  $\gamma = 0$  and  $(\alpha, \beta) \in \{(2, 0), (1, 1), (0, 1)\}$ , which gives three solutions:  $m \in \{44, 66, 33\}$ .

Further, let us have  $\delta = 0$ . If  $\gamma = 2$ , then  $\varphi(2^\alpha 3^\beta) = 1$ , whence  $(\alpha, \beta) \in \{(1, 0), (0, 0)\}$ . We thus obtain two more solutions:  $m \in \{50, 25\}$ .

If  $\gamma = 1$ , then we get  $\frac{20}{\varphi(5)} = 5$ , similarly to the above item. This is an odd integer, so we get no solutions in this case. This is also the case for  $\gamma = 0$  since the equation  $\varphi(2^\alpha) = 20$  has no solution, either.

Altogether, there are five satisfactory values  $m \in \{25, 33, 44, 50, 66\}$ .

- iv) This problem is of a different kind than the previous ones, so we must approach otherwise. The relation  $\varphi(m) = \frac{m}{3}$  implies that  $m$  must be a multiple of three (since the left-hand side of the equation is an integer). We will thus be looking for the solution in the form  $m = 3^\alpha \cdot n$ , where  $3 \nmid n, \alpha \geq 1$ . Then,  $\varphi(m) = 2 \cdot 3^{\alpha-1} \cdot \varphi(n) = \frac{m}{3} =$

**11.3.6. Dirichlet convolution.** There is another concept which is tightly connected to the Möbius function, the so-called Dirichlet product (also Dirichlet convolution).

**Definition.** Let  $f, g$  be arithmetic functions. Its *Dirichlet product* is defined as follows:

$$(f \circ g)(n) = \sum_{d \mid n} f(d) \cdot g\left(\frac{n}{d}\right) = \sum_{d_1 d_2 = n} f(d_1) \cdot g(d_2).$$

The next observation is nearly obvious:

**Lemma.** *The Dirichlet product is associative.*

PROOF.

$$((f \circ g) \circ h)(n) = \sum_{d_1 d_2 d_3 = n} f(d_1) g(d_2) h(d_3) = (f \circ (g \circ h))(n)$$

□

**11.3.7. Möbius inversion formula.** Let us define two useful functions  $i$  and  $I$  by  $i(1) = 1, i(n) = 0$  for all  $n > 1$  and  $I(n) = 1$  for all  $n \in \mathbb{N}$ . Then, every arithmetic function  $f$  satisfies:

$$f \circ i = i \circ f = f$$

$$(I \circ f)(n) = (f \circ I)(n) = \sum_{d \mid n} f(d).$$

Further, notice the Möbius function commutes with  $I$  and the result for all  $n > 1$  is  $i$  (the statement is clear for  $n = 1$  and we use the above Lemma 11.3.5):

$$(I \circ \mu)(n) = \sum_{d \mid n} I(d) \mu\left(\frac{n}{d}\right) = \sum_{d \mid n} I\left(\frac{n}{d}\right) \mu(d) =$$

$$= \sum_{d \mid n} \mu(d) = 0$$

**Theorem (Möbius inversion formula).** *Let an arithmetic function  $F$  be defined in terms of an arithmetic function  $f$  by  $F(n) = \sum_{d \mid n} f(d)$ . Then,  $f$  can be expressed as*

$$f(n) = \sum_{d \mid n} \mu\left(\frac{n}{d}\right) \cdot F(d).$$

PROOF. The relation  $F(n) = \sum_{d \mid n} f(d)$  can be rewritten as  $F = f \circ I$ . Therefore,  $F \circ \mu = (f \circ I) \circ \mu = f \circ (I \circ \mu) = f \circ i = f$ , which is the statement of the theorem. □

**11.3.8. Proof of the Theorem 11.3.3.** Invoking the Lemma 11.3.2 and the Möbius inversion formula, we get

$$\varphi(n) = \sum_{d \mid n} \mu(d) \frac{n}{d}$$

$$= n - \frac{n}{p_1} - \dots - \frac{n}{p_k} + \dots + (-1)^k \frac{n}{p_1 \cdots p_k}$$

$$= n \cdot \left(1 - \frac{1}{p_1}\right) \cdots \left(1 - \frac{1}{p_k}\right)$$

and the explicit formula for  $\varphi(n)$  has been proved.



ABBBB, AABBB, AAABB, AAAAB, ABABB, AABAB. □

**11.C.16.**

- i) Determine the remainder of the integer  $2^{50} + 3^{50} + 4^{50}$  when divided by 17.
- ii) Determine the remainder of the integer  $2^{181} + 3^{181} + 5^{181}$  when divided by 37.

**Solution.**

- i) By Fermat's theorem, we have  $2^{16} \equiv 3^{16} \equiv 4^{16} \equiv 1 \pmod{17}$ . Since  $50 \equiv 2 \pmod{16}$ , we get  $2^{50} + 3^{50} + 4^{50} \equiv 2^2 + 3^2 + 4^2 \equiv 12 \pmod{17}$ .
- ii) Similarly  $2^{36} \equiv 3^{36} \equiv 5^{36} \equiv 1 \pmod{37}$ , and hence  $2^{181} + 3^{181} + 5^{181} \equiv 2 + 3 + 5 \equiv 10 \pmod{37}$ . □

**Euler's theorem and orders of integers modulo  $m$ .** Thanks to Euler's theorem, it is guaranteed that every integer  $a$  which is coprime to  $m$  has an order, i.e. the least natural number  $n$  such that  $a^n \equiv 1 \pmod{m}$ . The most interesting ones are those integers  $a$  whose order equals  $\varphi(m)$ ; they are called the primitive roots modulo  $m$ .

**11.C.17.** Determine the order of 2 modulo 7.



**Solution.** The order of 2 modulo 7 is equal to 3 as  $2^1 = 2 \not\equiv 1 \pmod{7}$ ,  $2^2 = 4 \not\equiv 1 \pmod{7}$ ,  $2^3 = 8 \equiv 1 \pmod{7}$ . □

**11.C.18.** Determine the last two digits of the number  $7^{2019}$ .

**Solution.** We can easily see that the order of 7 modulo 100 is equal to 4 – by simple calculations, we have  $7^2 = 49$  and  $49^2 = (50 - 1)^2 = 50^2 - 2 \cdot 50 + 1 \equiv 1 \pmod{100}$ . Therefore, it suffices to determine the remainder  $r$  of the integer 2019 when divided by 4, since  $7^{2019} \equiv 7^r \pmod{100}$ . Apparently, we have  $r = 3$ , so the wanted last two digits are the same as those of  $7 \cdot 49$ , i.e. 43. □

Now, we mention several statements about the properties of the order of an integer modulo  $m$ .

**11.C.19.** Let  $m \in \mathbb{N}$ ,  $a, b \in \mathbb{Z}$ ,  $(a, m) = (b, m) = 1$ . Prove that if  $a \equiv b \pmod{m}$ , then the integers  $a, b$  share the same order modulo  $m$ .

**Solution.** Raising the congruence  $a \equiv b \pmod{m}$  to the  $n$ -th power leads to  $a^n \equiv b^n \pmod{m}$ , so  $a^n \equiv 1 \pmod{m} \iff b^n \equiv 1 \pmod{m}$ . □

**11.C.20.** Let  $m \in \mathbb{N}$ ,  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ . If the order of  $a$  modulo  $m$  is  $r \cdot s$ , (where  $r, s \in \mathbb{N}$ ), prove that the order of the integer  $a^r$  modulo  $m$  is  $s$ .

**Solution.** Since none of the integers  $a, a^2, a^3, \dots, a^{rs-1}$  is congruent to 1 modulo  $m$ , nor is any of the integers  $a^r, a^{2r}, a^{3r}, \dots, a^{(s-1)r}$ . On the other hand, we have  $(a^r)^s \equiv 1 \pmod{m}$ , so the order of  $a^r$  modulo  $m$  equals  $s$ . □

This formula can be generalized in many ways – especially in cases where the functions  $F$  and  $f$  map natural numbers into an abelian group  $(G, \cdot)$ . In this case, the formula (considering the operation in  $G$  to be multiplicative) gets the form

$$f(n) = \sum_{d|n} F(d) \mu\left(\frac{n}{d}\right).$$

Now, we will demonstrate the use of the Möbius inversion formula on a more complex example from the theory of finite fields. Let us consider a  $p$ -element field  $\mathbb{F}_p$  (i.e., the ring of residue classes modulo a prime  $p$ ) and examine the number  $N_d$  of monic irreducible polynomials of a given degree  $d$  over this field. Let  $S_d(x)$  denote the product of all such polynomials. Now, we borrow a (not very hard) theorem from the theory of finite fields which states that for all  $n \in \mathbb{N}$ , we have

$$x^{p^n} - x = \prod_{d|n} S_d(x).$$

Confronting the degrees of the polynomial on both sides yields

$$p^n = \sum_{d|n} dN_d,$$

whence we get, by applying the standard Möbius inversion formula, that

$$N_n = \frac{1}{n} \sum_{d|n} \mu\left(\frac{n}{d}\right) p^d.$$

In particular, we can see that for any  $n \in \mathbb{N}$ , it holds that  $N_n = \frac{1}{n} (p^n - \dots + \mu(n)p) \neq 0$  since the expression in the parentheses is a sum of distinct powers of  $p$  multiplied by coefficients  $\pm 1$ , so it cannot be equal to 0. Therefore, there exist irreducible polynomials over  $\mathbb{F}_p$  of an arbitrary degree  $n$ , so there are finite fields  $\mathbb{F}_{p^n}$  (having  $p^n$  elements) for any prime  $p$  and natural number  $n$  (in the theory of field extensions, such a field is constructed as the quotient ring  $\mathbb{F}_p[x]/(f)$  of the ring of polynomials over  $\mathbb{F}_p$  modulo the ideal generated by an irreducible polynomial  $f \in \mathbb{F}_p[x]$  of degree  $n$ , whose existence has just been proved).

**11.3.10. Example.** By the formula we have proved, the number of (monic) irreducible polynomials over  $\mathbb{F}_2$  of degree 5 is equal to

$$N_5 = \frac{1}{5} \sum_{d|5} \mu\left(\frac{5}{d}\right) 2^d = \frac{1}{5} (\mu(1) \cdot 2^5 + \mu(5) \cdot 2) = 6.$$

The number of monic irreducible polynomials over  $\mathbb{F}_3$  of degree four is then

$$N_4 = \frac{1}{4} \sum_{d|4} \mu\left(\frac{4}{d}\right) 3^d = \frac{1}{4} (\mu(1) \cdot 3^4 + \mu(2) \cdot 3^2 + \mu(4) \cdot 3^1) = \frac{1}{4} (81 - 9) = 18.$$

**11.3.11.** The next two theorems belong to the most important results of elementary number theory, and they will often be applied in further theoretical as well as practical problems.



**11.C.21.** Show that the contrary of the previous statement need not be true in general.

**Solution.** Indeed, even if the order of an integer  $a^r$  modulo  $m$  is  $s$ , the order of  $a$  modulo  $m$  may not be  $r \cdot s$ . For instance, for  $m = 13$  and the integers  $a = 3$ ,  $b = -4$ , we have  $a^2 = 9$ ,  $a^3 = 27 \equiv 1 \pmod{13}$ , so the order of  $a$  modulo 13 is 3. Similarly,  $b^2 = 16 \not\equiv 1 \pmod{13}$ ,  $b^3 = -64 \equiv 1 \pmod{13}$ , so the order of  $b$  modulo 13 is 3, too. On the other hand,  $b^2 = (-4)^2 = 16 \equiv 3 = a \pmod{13}$  has the same order (3) as  $a$ , yet the integer  $b$  does not have order  $2 \cdot 3$ .  $\square$

**11.C.22.** Determine the last digit of the numbers

- i)  $3^{5^{7^9}}$ ,
- ii)  $37^{37^{37}}$ ,
- iii)  $12^{13^{14}}$ .

**11.C.23.** It holds for all odd  $n \in \mathbb{N}$  that  $n \mid 2^{n!} - 1$ . Prove this!  $\circ$

**11.C.24.**

- i) Determine the last digit of the number  $7^{9^{5^{7^3}}}$ .
- ii) Determine the remainder of the number  $15^{14^{13}}$  when divided by 11.  $\circ$

**Solution.**

- i) The order of 7 modulo 100 is equal to 4 by exercise 11.C.18, so it suffices to find the remainder of the (extremely large) exponent upon division by 4. Since  $9 \equiv 1 \pmod{4}$ , the entire exponent leaves remainder 1 as well. Therefore, the wanted digit is  $7^1 = 7$ .
- ii) The order of  $15 \equiv 4 \pmod{11}$  is 5 (which can be found by direct computation or from the fact that 2 is a primitive root modulo 11 (see also 11.C.28); then, theorem 11.3.14 yields that the order of  $4 = 2^2$  is  $\frac{10}{(10,2)} = 5$ ). It is thus sufficient to determine the remainder of the exponent modulo 5. We have

$$14^{13} \equiv (-1)^{13} = -1 \equiv 4 \pmod{5},$$

so the wanted remainder is  $4^4 = 2^8 = 256 \equiv 6 - 5 + 2 = 3 \pmod{11}$ . Alternatively, we could have finished the calculation as follows:  $4^4 \equiv 4^{-1} \equiv 3 \pmod{11}$ .

**11.C.25.** Determine the last two digits of the decimal expansion of the number  $14^{14^{14}}$ .



**Solution.** We are interested in the remainder of the number  $a = 14^{14^{14}}$  upon division by 100. However, since  $(14, 100) > 1$ , we cannot consider the order of 14 modulo 100. Instead, we can factor the modulus to coprime integers:  $100 = 4 \cdot 25$ . Apparently,  $4 \mid a$ , so it remains to find the remainder of  $a$  modulo 25. By Euler's theorem, we have

$$14^{\varphi(25)} = 14^{20} \equiv 1 \pmod{25},$$

so we are interested in the remainder of  $14^{14}$  upon division by  $20 = 4 \cdot 5$ . Again, we clearly have  $4 \mid 14^{14}$ , and further  $14^{14} \equiv (-1)^{14} = 1 \pmod{5}$ , so

$$14^{14} \equiv 16 \pmod{20}.$$

Altogether,

$$14^{14^{14}} \equiv 14^{16} = 2^{16} \cdot 7^{16} \pmod{25}.$$

FERMAT'S LITTLE THEOREM

**Theorem.** Let  $a$  be an integer and  $p$  a prime,  $p \nmid a$ . Then,

$$a^{p-1} \equiv 1 \pmod{p}.$$

**PROOF.** The statement will follow as a simple consequence of Euler's theorem (and together with this one, it is a consequence of more general Lagrange's theorem 12.4.10). However, it can be proved directly (by mathematical induction or a combinatorial means, as mentioned in exercise 11.C.15).  $\square$

Sometimes, Fermat's little theorem is presented in the following form, which is apparently equivalent to the original statement.

**Corollary.** Let  $a$  be an integer and  $p$  a prime. Then,

$$a^p \equiv a \pmod{p}.$$

**11.3.12. Euler's theorem.** We are coming to a theorem which has got immediate striking applications in cryptography (we shall come to this point in cf. 11.6.18). Before formulating and proving Euler's theorem, we introduce a few useful concepts.

RESIDUE SYSTEMS

A *complete residue system* modulo  $m$  is an arbitrary  $m$ -tuple of integers which are pairwise incongruent modulo  $m$  (the most commonly used  $m$ -tuple is  $0, 1, \dots, m-1$  or, for odd  $m$ , its "symmetric" variation  $-\frac{m-1}{2}, \dots, -1, 0, 1, \dots, \frac{m-1}{2}$ ).

A *reduced residue system* modulo  $m$  is an arbitrary  $\varphi(m)$ -tuple of integers which are pairwise incongruent modulo  $m$  and coprime to  $m$ .

**Lemma.** Let  $x_1, x_2, \dots, x_{\varphi(m)}$  form a reduced residue system modulo  $m$ . If  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ , then the integers  $a \cdot x_1, \dots, a \cdot x_{\varphi(m)}$  also form a reduced residue system modulo  $m$ .

**PROOF.** Since  $(a, m) = 1$  and  $(x_i, m) = 1$ , we have  $(a \cdot x_i, m) = 1$ . Further, if we had  $a \cdot x_i \equiv a \cdot x_j \pmod{m}$  for some distinct indices  $i, j$ , dividing both sides of the congruence by the integer  $a$  (which is coprime to  $m$ ) would lead to  $x_i \equiv x_j \pmod{m}$ , meaning that the original  $\varphi(m)$ -tuple was not a reduced residue system, either.  $\square$

EULER'S THEOREM

**Theorem.** Let  $a \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ ,  $(a, m) = 1$ . Then,

$$a^{\varphi(m)} \equiv 1 \pmod{m}.$$

**PROOF.** Let  $x_1, x_2, \dots, x_{\varphi(m)}$  be an arbitrary reduced residue system modulo  $m$ . By the previous lemma,  $a \cdot x_1, \dots, a \cdot x_{\varphi(m)}$  is also a reduced residue system modulo  $m$ .

We can simplify the computation to come a lot if we realize that

$$7^2 \equiv -1 \pmod{25}, \text{ and } 2^5 \equiv 7 \pmod{25}.$$

Then,

$$\begin{aligned} 14^{14^{14}} &\equiv 2^{16} \cdot 7^{16} \equiv (2^5)^3 \cdot 2 \cdot 7^{16} \equiv 7^3 \cdot 2 \cdot 7^{16} \\ &\equiv 2 \cdot 7^{19} \equiv 2 \cdot (-1)^9 \cdot 7 = 11 \pmod{25}. \end{aligned}$$

We are thus looking for a non-negative integer which is less than 100, is a multiple of 4, and leaves remainder 11 when divided 25 – the only such number is clearly 36.  $\square$

**11.C.26.** Determine the last three digits of the number  $12^{10^{11}}$ .  $\circ$

**11.C.27.** Find all natural numbers  $n$  for which the integer  $5^n - 4^n - 3^n$  is divisible by eleven.



**Solution.** The orders of all of the numbers 3, 4, and 5 are equal to five, so it suffices to examine  $n \in \{0, 1, 2, 3, 4\}$ . It can be seen from the following table

$n$	0	1	2	3	4
$5^n \pmod{11}$	1	5	3	4	9
$4^n \pmod{11}$	1	4	5	9	3
$3^n \pmod{11}$	1	3	9	5	4

that only the case  $n \equiv 2 \pmod{5}$  yields  $3 - 5 - 9 \equiv 0 \pmod{11}$ .

The problem is thus satisfied by exactly those natural numbers  $n$  which satisfy  $n \equiv 2 \pmod{5}$ .  $\square$

**11.C.28. Primitive roots.** Show that there are no primitive roots modulo 8, and find any primitive root modulo 11.

**Solution.** Apparently, even integers cannot be primitive roots modulo 8, so it remains to examine odd ones. We can easily calculate that  $3^2 \equiv 5^2 \equiv 7^2 \equiv 1 \pmod{8}$ , but  $\varphi(8) = 4 > 2$ . Now, we will verify that 2 is a primitive root modulo 11. The order of 2 divides  $\varphi(11) = 10$ , so it suffices to show that  $2^2 \not\equiv 1 \pmod{11}$  and  $2^5 = 32 \equiv -1 \not\equiv 1 \pmod{11}$ . Therefore, the order of 2 modulo 11 is indeed 10.  $\square$

**11.C.29.** We will now determine (with the help of propositions proving Theorem 11.3.16) primitive roots modulo 41,  $41^2$ , and  $2 \cdot 41^2$ .



**Solution.** Since  $\varphi(41) = 40 = 2^3 \cdot 5$ , it holds that an integer  $g$  coprime to 41 is a primitive root modulo 41 if and only if

$$g^{20} \not\equiv 1 \pmod{41} \wedge g^8 \not\equiv 1 \pmod{41}.$$

Therefore, for every  $i \in \{1, 2, \dots, \varphi(m)\}$ , there is a unique  $j \in \{1, 2, \dots, \varphi(m)\}$  such that  $a \cdot x_i \equiv x_j \pmod{m}$ . Multiplying these congruences leads to  $(a \cdot x_1) \cdot (a \cdot x_2) \cdots (a \cdot x_{\varphi(m)}) \equiv x_1 \cdot x_2 \cdots x_{\varphi(m)} \pmod{m}$ , which can be rearranged to

$$a^{\varphi(m)} \cdot x_1 \cdot x_2 \cdots x_{\varphi(m)} \equiv x_1 \cdot x_2 \cdots x_{\varphi(m)} \pmod{m}.$$

Dividing this by the integer  $x_1 \cdot x_2 \cdots x_{\varphi(m)}$  already gives the wanted statement.  $\square$

**Remark.** As we have already mentioned, Euler's theorem is a consequence of Lagrange's theorem (see 12.4.10) applied to the group  $(\mathbb{Z}_m^\times, \cdot)$ . This proof of Euler's theorem utilized the fact that multiplying by an integer  $a$  which is coprime to  $m$  is, in algebraic words, an automorphism of the group  $(\mathbb{Z}_m^\times, \cdot)$ .

**11.3.13.** There is an important concept which is tightly connected to Euler's totient function and Euler's theorem: the so-called *order of an integer modulo  $m$*  – once again, it is nothing else than the order of the corresponding element in the group of invertible residue classes modulo  $m$ :

ORDER OF AN INTEGER

Let  $a \in \mathbb{Z}$ ,  $m \in \mathbb{N}$ , where  $(a, m) = 1$ . The *order of  $a$  modulo  $m$*  is the least natural number  $n$  satisfying

$$a^n \equiv 1 \pmod{m}.$$

It follows from Euler's theorem that the order of an integer is well-defined – the order of any integer coprime to the modulus is surely not greater than  $\varphi(m)$ . As we will see later, the integers whose order is exactly  $\varphi(m)$  are of great interest – they are called *primitive roots* modulo  $m$  and they play an important role in solving binomial congruences, among others. This concept is just another name for a generator of the group  $(\mathbb{Z}_m^\times, \cdot)$ .

Some of the very basic results of the order are demonstrated in 11.C.19, and complete description of the dependency of the order upon the exponent is given by the subsequent two theorems.

**Theorem.** Let  $m \in \mathbb{N}$ ,  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ . Let  $r$  denote the order of  $a$  modulo  $m$ . Then, for any  $t, s \in \mathbb{N}_0$ , we have

$$a^t \equiv a^s \pmod{m} \iff t \equiv s \pmod{r}.$$

**PROOF.** Without loss of generality, we can assume that  $t \geq s$ . Dividing the integer  $t - s$  by  $r$  with remainder, we get  $t - s = q \cdot r + z$ , where  $q, z \in \mathbb{N}_0, 0 \leq z < r$ .

“ $\Leftarrow$ ” Since  $t \equiv s \pmod{r}$ , we have  $z = 0$ , hence  $a^{t-s} = a^{qr} = (a^r)^q \equiv 1^q \pmod{m}$ . Multiplying both sides of the congruence by the integer  $a^s$  leads to the wanted statement.

“ $\Rightarrow$ ” It follows from  $a^t \equiv a^s \pmod{m}$  that  $a^s \cdot a^{qr+z} \equiv a^s \pmod{m}$ . Since  $a^r \equiv 1 \pmod{m}$ , we also have  $a^{qr+z} \equiv a^z \pmod{m}$ . Altogether, after dividing both sides of the first congruence by the integer  $a^s$  (which is coprime to the modulus), we get  $a^z \equiv 1 \pmod{m}$ . Since



Now, we will go through the potential primitive roots in ascending order:

$$g = 2 : \quad 2^8 = 2^5 \cdot 2^3 \equiv -9 \cdot 8 \equiv 10 \pmod{41},$$

$$2^{20} = (2^5)^4 \equiv (-9)^4 = 81^2 \equiv (-1)^2$$

$$\equiv 1 \pmod{41},$$

$$g = 3 : \quad 3^8 = (3^4)^2 \equiv (-1)^2 = 1 \pmod{41},$$

$g = 4$  : the order of 4 = 2<sup>2</sup> always divides the order 2,

$$g = 5 : \quad 5^8 = (5^2)^4 \equiv (-2^4)^4 = 2^{16} = (2^8)^2$$

$$\equiv 10^2 \equiv 18 \pmod{41},$$

$$5^{20} = (5^2)^{10} \equiv (-2^4)^{10} = 2^{40} = (2^{20})^2$$

$$\equiv 1 \pmod{41},$$

$$g = 6 : \quad 6^8 = 2^8 \cdot 3^8 \equiv 10 \cdot 1 = 10 \pmod{41},$$

$$6^{20} = 2^{20} \cdot 3^{20} \equiv 2^{20} \cdot (3^8)^2 \cdot 3^4$$

$$\equiv 1 \cdot 1 \cdot (-1) = -1 \pmod{41}.$$

We have thus proved that 6 is the least positive primitive root modulo 41 (if we were interested in other primitive roots modulo 41 as well, we would get them as the powers of 6 with exponent taking on values from the range 1 to 40 which are coprime to 40. There are exactly  $\varphi(40) = \varphi(2^3 \cdot 5) = 16$  of them, and the resulting remainders modulo 41 are  $\pm 6, \pm 7, \pm 11, \pm 12, \pm 13, \pm 15, \pm 17, \pm 19$ ).

Now, if we prove that  $6^{40} \not\equiv 1 \pmod{41^2}$ , we will know that 6 is a primitive root modulo any power of 41 (if we had “bad luck” and found out that  $6^{40} \equiv 1 \pmod{41^2}$ , then a primitive root modulo  $41^2$  would be  $47 = 6 + 41$ ). To avoid manipulating huge numbers when verifying the condition, we will use several tricks (the so-called residue number system).

First of all, we calculate the remainder of  $6^8$  upon division by  $41^2$ ; this problem can be further reduced to computing the remainders of the integers  $2^8$  and  $3^8$ :

$$2^8 = 256 = 6 \cdot 41 + 10 \pmod{41^2},$$

$$3^8 = (3^4)^2 = (2 \cdot 41 - 1)^2 \equiv -4 \cdot 41 + 1 \pmod{41^2}.$$

Then,

$$6^8 = 2^8 \cdot 3^8 \equiv (6 \cdot 41 + 10)(-4 \cdot 41 + 1)$$

$$\equiv -34 \cdot 41 + 10 \equiv 7 \cdot 41 + 10 \pmod{41^2}$$

and

$$6^{40} = (6^8)^5 \equiv (7 \cdot 41 + 10)^5 \equiv (10^5 + 5 \cdot 7 \cdot 41 \cdot 10^4)$$

$$= 10^4(10 + 35 \cdot 41) \equiv (-2 \cdot 41 - 4)(-6 \cdot 41 + 10)$$

$$\equiv (4 \cdot 41 - 40) = 124 \not\equiv 1 \pmod{41^2}.$$

In the calculation, we made use of the fact that  $10^4 = 6 \cdot 41^2 - 86$ , i.e.,  $10^4 \equiv -2 \cdot 41 - 4 \pmod{41^2}$ .

Therefore, 6 is a primitive root modulo  $41^2$ , and since it is an even integer, we can see that  $1687 = 6 + 41^2$  is a primitive root modulo  $2 \cdot 41^2$  (while the least positive primitive root modulo  $2 \cdot 41^2$  is the integer 7).  $\square$

$z < r$ , it follows from the definition of the order that  $z = 0$ , hence  $r \mid t - s$ .  $\square$

The above theorem and Euler’s theorem apparently lead to the following corollary (whose second part is only a reformulation of Lagrange’s theorem 12.4.10 for our situation):

**Corollary.** Let  $m \in \mathbb{N}$ ,  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ , and  $r$  be the order of  $a$  modulo  $m$ .

(1) For any  $n \in \mathbb{N} \cup \{0\}$ , it holds that

$$a^n \equiv 1 \pmod{m} \iff r \mid n.$$

(2)  $r \mid \varphi(m)$

**11.3.14.** The following theorem is a generalization of the result in 11.C.20.

**Theorem.** Let  $m, n \in \mathbb{N}$ ,  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ . If the order of  $a$  modulo  $m$  is  $r$ , then the order of  $a^n$  modulo  $m$  is  $\frac{r}{(r,n)}$ .

**PROOF.** Since  $\frac{r \cdot n}{(r,n)} = [r, n]$ , which is clearly a multiple of  $r$ , we have

$$(a^n)^{\frac{r}{(r,n)}} = a^{[r,n]} \equiv 1 \pmod{m}$$

(the last statement follows from the above corollary, because  $r \mid [r, n]$ ). On the other hand, if  $k \in \mathbb{N}$  is such that  $(a^n)^k = a^{n \cdot k} \equiv 1 \pmod{m}$ , we get  $r \mid n \cdot k$  (since  $r$  is the order of  $a$ ). Further, we know that  $\frac{r}{(n,r)} \mid \frac{n}{(n,r)} \cdot k$ , whence (thanks to  $\frac{r}{(n,r)}$  and  $\frac{n}{(n,r)}$  being coprime)  $\frac{r}{(n,r)} \mid k$ . Therefore,  $\frac{r}{(n,r)}$  is the order of the integer  $a^n$  modulo  $m$ .  $\square$

The last statement of this series connects the orders of two integers to the order of their product:

**Lemma.** Let  $m \in \mathbb{N}$ ,  $a, b \in \mathbb{Z}$ ,  $(a, m) = (b, m) = 1$ . If  $a$  has order  $r$  and  $b$  has order  $s$  modulo  $m$ , where  $(r, s) = 1$ , then the integer  $a \cdot b$  has order  $r \cdot s$  modulo  $m$ .

**PROOF.** Let  $\delta$  denote the order of  $a \cdot b$ . Then,  $(ab)^\delta \equiv 1 \pmod{m}$ . Raising both sides of this congruence to the  $r$ -th power leads to  $a^{r\delta} b^{r\delta} \equiv 1 \pmod{m}$ . Since  $r$  is the order of  $a$ , we have  $a^r \equiv 1 \pmod{m}$ , i.e.,  $b^{r\delta} \equiv 1 \pmod{m}$ , and so  $s \mid r\delta$ . From  $r$  being coprime to  $s$ , we get  $s \mid \delta$ . Analogously, we can get  $r \mid \delta$ , so (again utilizing that  $r, s$  are coprime)  $r \cdot s \mid \delta$ . On the other hand, we clearly have  $(ab)^{rs} \equiv 1 \pmod{m}$ , hence  $\delta \mid rs$ . Altogether,  $\delta = rs$ .  $\square$

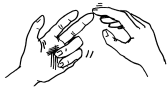
**11.3.15. Primitive roots.** Among the integers coprime to a modulus  $m$  (i.e., the elements of a reduced residue system modulo  $m$ ), the most important ones are those whose order is equal to  $\varphi(m)$ .



Step-by-step exponentiation of such a number yields all possible elements of a reduced residue system (or integers congruent to them). Therefore, in various problems, we can work with powers of a given integer instead of considering random elements of a reduced residue system modulo  $m$ , and this is often much simpler (see, for instance, the proof of the theorem 11.4.10 about binomial coefficients).

**D. Solving congruences**

**Linear congruences.** The following exercise illustrates that the procedure we mentioned in the proof of theorem 11.4.3 about solvability of linear congruences (which invokes Euler’s theorem) is usually not the most efficient one – we can utilize both Bézout’s theorem and equivalent modifications of the given congruence.



**11.D.1.** Solve the congruence

$$39x \equiv 41 \pmod{47}.$$

**Solution.**

- i) First, we use Euler’s theorem. Since  $(39, 47) = 1$ , we have

$$39^{\varphi(47)} = 39^{46} \equiv 1 \pmod{47},$$

i.e.,

$$\underbrace{39^{45} \cdot 39}_{39^{46} \equiv 1} x \equiv 39^{45} \cdot 41 \pmod{47},$$

whence it already follows that

$$x \equiv 39^{45} \cdot 41 \pmod{47}.$$

To complete the solution, it remains to calculate the remainder of  $39^{45} \cdot 41$  when divided by 47, which is left as an exercise to the kind reader, leading to the result  $x \equiv 36 \pmod{47}$ .

- ii) Another option is to make use of Bézout’s theorem. The Euclidean algorithm applied to the pair  $(39, 47)$  yields

$$\begin{aligned} 47 &= 1 \cdot 39 + 8, \\ 39 &= 4 \cdot 8 + 7, \\ 8 &= 1 \cdot 7 + 1. \end{aligned}$$

In the other direction, this leads to

$$\begin{aligned} 1 &= 8 - 7 = 8 - (39 - 4 \cdot 8) = 5 \cdot 8 - 39 \\ &= 5 \cdot (47 - 39) - 39 = 5 \cdot 47 - 6 \cdot 39. \end{aligned}$$

Considering this equality modulo 47 and remembering that we are solving the equation  $41 \equiv x \cdot 39$ , we obtain

$$\begin{aligned} 1 &\equiv -6 \cdot 39 && \pmod{47}, && / \cdot 41 \\ 41 &\equiv \underbrace{41 \cdot (-6)} \cdot 39 && \pmod{47}, \\ x &\equiv 41 \cdot (-6) && \pmod{47}, \\ x &\equiv -246 && \pmod{47}, \\ x &\equiv 36 && \pmod{47}. \end{aligned}$$

Let us notice that this procedure is usually used in the corresponding software tools – it is efficient and can be easily made into an algorithm. It was also important that 41 (the number we multiplied the congruence with) and the modulus 47 are coprime.

PRIMITIVE ROOT

Let  $m \in \mathbb{N}$ . An integer  $g$ ,  $(g, m) = 1$ , is said to be a *primitive root modulo  $m$*  if and only if its order modulo  $m$  equals  $\varphi(m)$ .

**Lemma.** *If  $g$  is a primitive root modulo  $m$ , then for every integer  $a$  such that  $(a, m) = 1$ , there is a unique  $x_a \in \mathbb{Z}$ ,  $0 \leq x_a < \varphi(m)$  with the property that  $g^{x_a} \equiv a \pmod{m}$ .*

The mapping  $a \mapsto x_a$  is called the *discrete logarithm* or *index* of the integer  $a$  (with respect to a given modulus  $m$  and a fixed primitive root  $g$ ), and it is a bijection between the sets  $\{a \in \mathbb{Z}; (a, m) = 1, 0 < a < m\}$  and  $\{x \in \mathbb{Z}; 0 \leq x < \varphi(m)\}$ .

**PROOF.** Suppose that it holds for  $x, y \in \mathbb{Z}$ ,  $0 \leq x, y < \varphi(m)$  that  $g^x \equiv g^y \pmod{m}$ . From the properties of the order, we get  $x \equiv y \pmod{\varphi(m)}$ , i.e.,  $x = y$ , so the mapping is injective. Since it is a mapping between two finite sets which have the same number of elements, it must be surjective as well.  $\square$

If there are primitive roots at all for a natural number  $m$ , then there are exactly  $\varphi(\varphi(m))$  of them among the integers  $1, 2, \dots, m$ : If  $g$  is a primitive root and  $a \in \{1, 2, \dots, \varphi(m)\}$  arbitrary, then the order of  $g^a$  is  $\frac{\varphi(m)}{(a, \varphi(m))}$  (by theorem 11.3.14), which is equal to  $\varphi(m)$  if and only if  $(a, \varphi(m)) = 1$ , and there are exactly  $\varphi(\varphi(m))$  such integers in the set  $\{1, 2, \dots, \varphi(m)\}$ .

Now, we are about to show that primitive roots exist for a sufficient amount of moduli  $m$ .

**11.3.16. Theorem** (Existence of primitive roots). *Let  $m \in \mathbb{N}$ ,  $m > 1$ . The modulus  $m$  has primitive roots if and only if at least one of the following conditions holds:*

- $m = 2$  or  $m = 4$ ,
- $m$  is a power of an odd prime,
- $m$  is twice a power of an odd prime.

The proof of this theorem will be done in several steps. We can easily see that 1 is a primitive root modulo 2 and 3 is a primitive root modulo 4. Further, we will show that primitive roots exist modulo any odd prime (in algebraic words, this is another proof of the fact that the group  $(\mathbb{Z}_m^\times, \cdot)$  of invertible residue classes modulo a prime  $m$  is cyclic; see also 12.4.8).

**Proposition.** *Let  $p$  be an odd prime. Then there are primitive roots modulo  $p$ .*

**PROOF.** Let  $r_1, r_2, \dots, r_{p-1}$  be the orders of the integers  $1, 2, \dots, p - 1$  modulo  $p$ . Let  $\delta = [r_1, r_2, \dots, r_{p-1}]$  be the least common multiple of these orders. We will show that there is an integer of order  $\delta$  among  $1, 2, \dots, p - 1$  and that  $\delta = p - 1$ .

Let  $\delta = q_1^{\alpha_1} \cdots q_k^{\alpha_k}$  be the factorization of  $\delta$  to primes. For every  $s \in \{1, \dots, k\}$ , there is a  $c \in \{1, \dots, p - 1\}$  such that  $q_s^{\alpha_s} \mid r_c$  (otherwise, there would be a common multiple



iii) Concerning paper-and-pencil calculations, the most efficient procedure (yet one not easily generalizable into an algorithm) is to gradually modify the congruence so that the set of solutions remains unchanged:

$$\begin{aligned} 39x &\equiv 41 && (\text{mod } 47), \\ -8x &\equiv -6 && (\text{mod } 47), && / : -2 \\ 4x &\equiv 3 && (\text{mod } 47), \\ 4x &\equiv -44 && (\text{mod } 47), && / : 4 \\ x &\equiv -11 && (\text{mod } 47), \\ x &\equiv 36 && (\text{mod } 47). \end{aligned}$$

□

**Systems of congruences.** In order to solve system of (not only linear) congruences, we will often utilize the Chinese remainder theorem, which guarantees uniqueness of the solution provided the moduli of the particular congruences are pairwise coprime.

**11.D.2.** Solve the system

$$\begin{aligned} x &\equiv 7 \pmod{27}, \\ x &\equiv -3 \pmod{11}. \end{aligned}$$

**Solution.** As  $(27, 11) = 1$ , we are guaranteed by the Chinese remainder theorem that the solution is unique modulo  $27 \cdot 11 = 297$ . There are two major possible approaches to finding the solution.

(a) Using the Euclidean algorithm, we can find the coefficients in Bézout's identity:  $1 = 5 \cdot 11 - 2 \cdot 27$ . Hence,  $[11]_{27}^{-1} = [5]_{27}$  and  $[27]_{11}^{-1} = [-2]_{11}$ . Therefore, the solution is  $x \equiv 7 \cdot 11 \cdot 5 - 3 \cdot 27 \cdot (-2) = 547 \equiv 250 \pmod{297}$ .

(b) Using step-by-step substitution, we get  $x = 11t - 3$  from the second congruence. Substituting this into the first one leads to  $11t \equiv 10 \pmod{27}$ . Multiplying this by 5 yields  $55t \equiv 50$ , i.e.,  $t \equiv -4 \pmod{27}$ . Altogether,  $x = 11 \cdot 27 \cdot s - 4 \cdot 11 - 3 = 297s - 47$  for  $s \in \mathbb{Z}$ , i.e.,  $x \equiv -47 \pmod{297}$ . □

**11.D.3.** Solve the following system of congruences:

$$\begin{aligned} x &\equiv 1 \pmod{10}, \\ x &\equiv 5 \pmod{18}, \\ x &\equiv -4 \pmod{25}. \end{aligned}$$

**Solution.** The integers  $x$  which satisfy the first congruence are those of the form  $x = 1 + 10t$ , where  $t \in \mathbb{Z}$  may be arbitrary. We will substitute this expression into the second congruence and then solve it (as a congruence in variable  $t$ ):

$$\begin{aligned} 1 + 10t &\equiv 5 \pmod{18}, \\ 10t &\equiv 4 \pmod{18}, \\ 5t &\equiv 2 \pmod{9}, \\ 5t &\equiv 20 \pmod{9}, \\ t &\equiv 4 \pmod{9}, \end{aligned}$$

of the integers  $r_1, r_2, \dots, r_{p-1}$  less than  $\delta$ ). Therefore, there exists an integer  $b$  such that  $r_c = b \cdot q_s^{\alpha_s}$ . Since  $c$  has order  $r_c$ , the order of the integer  $g_s := c^b$  is equal to  $q_s^{\alpha_s}$  (by the theorem 11.3.14 on orders of powers).

Reasoning analogously for any  $s \in \{1, \dots, k\}$ , we get integers  $g_1, \dots, g_k$ , and we can set  $g := g_1 \cdots g_k$ . From the properties of the order of a product, we get that the order of  $g$  is equal to the product of the orders of the integers  $g_1, \dots, g_k$ , i.e. to  $q_1^{\alpha_1} \cdots q_k^{\alpha_k} = \delta$ .

Now, we prove that  $\delta = p - 1$ . Since the orders of the integers  $1, 2, \dots, p - 1$  divide  $\delta$ , we get the congruence  $x^\delta \equiv 1 \pmod{p}$  for any  $x \in \{1, 2, \dots, p - 1\}$ . By theorem 11.4.8, there are at most  $\delta$  solutions to a congruence of degree  $\delta$  modulo a prime  $p$  (in algebraic words, we are actually looking for roots of a polynomial over a field, and there cannot be more of them than the degree of the polynomial, as we will see in part 12.3.4). On the other hand, we have already shown that this congruence has  $p - 1$  solutions, so necessarily  $\delta \geq p - 1$ . Still,  $\delta$  is (being the order of  $g$ ) a divisor of  $p - 1$ , whence we finally get the wanted equality  $\delta = p - 1$ . □

**11.3.17.** Next, we show that there are primitive roots modulo powers of odd primes. First, we prove two helpful lemmas and then comes the proposition.

**Lemma.** Let  $p$  be an odd prime,  $\ell \geq 2$  arbitrary. Then, it holds for any  $a \in \mathbb{Z}$  that

$$(1 + ap)^{p^{\ell-2}} \equiv 1 + ap^{\ell-1} \pmod{p^\ell}.$$

**PROOF.** This will follow easily from the binomial theorem using mathematical induction on  $\ell$ .

I. The statement is clearly true for  $\ell = 2$ .

II. Let the statement be true for  $\ell$ , and let us prove it for  $\ell + 1$ . Invoking exercise 11.C.7 and raising the statement for  $\ell$  to the  $p$ -th power, we obtain

$$(1 + ap)^{p^{\ell-1}} \equiv (1 + ap^{\ell-1})^p \pmod{p^{\ell+1}}.$$

It follows from the binomial theorem that

$$(1 + ap^{\ell-1})^p = 1 + p \cdot a \cdot p^{\ell-1} + \sum_{k=2}^p \binom{p}{k} a^k p^{(\ell-1)k}$$

and since we have  $p \mid \binom{p}{k}$  for  $1 < k < p$  (by exercise 11.C.6), it suffices to show that  $p^{\ell+1} \mid p^{1+(\ell-1)k}$ , which is equivalent to  $1 \leq (k - 1)(\ell - 1)$ . Thanks to the assumption  $\ell \geq 2$ , we get that  $p^{\ell+1} \mid p^{(\ell-1)p}$  for  $k = p$  as well. □

**Lemma.** Let  $p$  be an odd prime,  $\ell \geq 2$  arbitrary. Then, it holds for any integer  $a$  satisfying  $p \nmid a$  that the order of  $1 + ap$  modulo  $p^\ell$  equals  $p^{\ell-1}$ .

**PROOF.** By the previous lemma, we have

$$(1 + ap)^{p^{\ell-1}} \equiv 1 + ap^\ell \pmod{p^{\ell+1}},$$

and considering this congruence modulo  $p^\ell$ , we get  $(1 + ap)^{p^{\ell-1}} \equiv 1 \pmod{p^\ell}$ . At the same time, it follows directly from the previous lemma and  $p$  not being a divisor of  $a$  that

or  $t = 4 + 9s$ , where  $s \in \mathbb{Z}$  is arbitrary. The first two congruences are thus satisfied by exactly those integers  $x$  which are of the form  $x = 1 + 10t = 1 + 10(4 + 9s) = 41 + 90s$ .

Once again, this can be substituted into the third congruence and then solved:

$$\begin{aligned} 41 + 90s &\equiv -4 \pmod{25}, \\ 90s &\equiv 5 \pmod{25}, \\ 18s &\equiv 1 \pmod{5}, \\ 3s &\equiv 6 \pmod{5}, \\ s &\equiv 2 \pmod{5}, \end{aligned}$$

or  $s = 2 + 5r$ , where  $r \in \mathbb{Z}$ . Altogether,  $x = 41 + 90s = 41 + 90(2 + 5r) = 221 + 450r$ .

Therefore, the system is satisfied by those integers  $x$  with  $x \equiv 221 \pmod{450}$ .  $\square$

**11.D.4.** A group of thirteen pirates managed to steal a chest full of gold coins (there were around two thousand of them). The pirates tried to divide them evenly among themselves, but ten coins were left over. They started to fight for the remaining coins, and one of the pirates was deadly stabbed during the combat. So, they tried to divide the coins evenly once again, and now three coins were left. Another pirate died in a subsequent battle for the three coins. The remaining pirates tried to divide the coins evenly for the third time, now successfully. How many coins were there in the chest?

**Solution.** The problem leads to the following system of congruences:

$$\begin{aligned} x &\equiv 10 \pmod{13}, \\ x &\equiv 3 \pmod{12}, \\ x &\equiv 0 \pmod{11}. \end{aligned}$$

Its solution is  $x \equiv 231 \pmod{11 \cdot 12 \cdot 13}$ . Since the number  $x$  of coins is to be around 2000 and  $x \equiv 231 \pmod{1716}$ , we can easily settle that there were exactly  $231 + 1716 = 1947$  coins.  $\square$

**11.D.5.** When gymnasts made groups of eight people, three were left over. When they formed circles, each consisting of seventeen people, seven remained; and when they grouped into pyramids (each of them contains  $21 = 4^2 + 2^2 + 1$  gymnasts), two of them were incomplete (missing a person “on the top”). How many gymnasts were there, provided there were at least 2000 and at most 4000?

**Solution.** We solve the following system of linear congruences in the standard way:

$$\begin{aligned} c &\equiv 3 \pmod{8}, \\ c &\equiv 7 \pmod{17}, \\ c &\equiv -2 \pmod{21}, \end{aligned}$$

$(1 + ap)^{p^{\ell-2}} \not\equiv 1 \pmod{p^\ell}$ , which gives the wanted proposition.  $\square$

**Proposition.** Let  $p$  be an odd prime. Then, for every  $\ell \in \mathbb{N}$ , there is a primitive root modulo  $p^\ell$ .

**PROOF.** Let  $g$  be a primitive root modulo  $p$ . We will show that if  $g^{p-1} \not\equiv 1 \pmod{p^2}$ , then  $g$  is a primitive root even modulo  $p^\ell$  for any  $\ell \in \mathbb{N}$ . (If we had  $g^{p-1} \equiv 1 \pmod{p^2}$ , then  $(g + p)^{p-1} \equiv 1 + (p-1)g^{p-2}p \not\equiv 1 \pmod{p^2}$ , so we could choose  $g + p$  for the original primitive root instead of the congruent integer  $g$ .)

Let  $g$  satisfy  $g^{p-1} \not\equiv 1 \pmod{p^2}$ . Then, there is an  $a \in \mathbb{Z}$ ,  $p \nmid a$  such that  $g^{p-1} = 1 + p \cdot a$ . We will show that the order of  $g$  modulo  $p^\ell$  is  $\varphi(p^\ell) = (p-1)p^{\ell-1}$ . Let  $n$  be the least natural number which satisfies  $g^n \equiv 1 \pmod{p^\ell}$ . By the previous lemma, the order of  $g^{p-1} = 1 + p \cdot a$  modulo  $p^\ell$  is  $p^{\ell-1}$ . However, then it follows from the corollary of 11.3.13 that

$$(g^{p-1})^n = (g^n)^{p-1} \equiv 1 \pmod{p^\ell} \implies p^{\ell-1} \mid n.$$

At the same time, the congruence  $g^n \equiv 1 \pmod{p}$  implies that  $p-1 \mid n$ . From  $p-1$  and  $p^{\ell-1}$  being coprime, we get that  $(p-1)p^{\ell-1} \mid n$ . Therefore,  $n = \varphi(p^\ell)$ , and  $g$  is thus a primitive root modulo  $p^\ell$ .  $\square$

**11.3.18.** Our next task is to deal with the existing primitive roots in form of multiples of two.

**Proposition.** Let  $p$  be an odd prime and  $g$  a primitive root modulo  $p^\ell$  for  $\ell \in \mathbb{N}$ . Then the odd one of the integers  $g, g+p^\ell$  is a primitive root modulo  $2p^\ell$ .

**PROOF.** Let  $c$  be an odd natural number. Then, for every  $n \in \mathbb{N}$ , we have  $c^n \equiv 1 \pmod{p^\ell}$  if and only if  $c^n \equiv 1 \pmod{2p^\ell}$ . Since  $\varphi(2p^\ell) = \varphi(p^\ell)$ , every odd primitive root modulo  $p^\ell$  is also a primitive root modulo  $2p^\ell$ .  $\square$

The subsequent proposition describes the case of powers of two. We will use similar helping lemmas as in the case of odd primes.

**Lemma.** Let  $\ell \in \mathbb{N}$ ,  $\ell \geq 3$ . Then  $5^{2^{\ell-3}} \equiv 1 + 2^{\ell-1} \pmod{2^\ell}$ .

**PROOF.** Similarly as above for  $2 \nmid p$ .  $\square$

**Lemma.** Let  $\ell \in \mathbb{N}$ ,  $\ell \geq 3$ . Then the order of the integer 5 modulo  $2^\ell$  is  $2^{\ell-2}$ .

**PROOF.** Easily from the above lemma.  $\square$

**Proposition.** Let  $\ell \in \mathbb{N}$ . There are primitive roots modulo  $2^\ell$  if and only if  $\ell \leq 2$ .

**PROOF.** Let  $\ell \geq 3$ . Then the set

$$S = \{(-1)^a \cdot 5^b; a \in \{0, 1\}, 0 \leq b < 2^{\ell-2}; b \in \mathbb{Z}\}$$

forms a reduce residue system modulo  $2^\ell$ : it has  $\varphi(2^\ell)$  elements, and it can be easily verified that they are pairwise incongruent modulo  $2^\ell$ .

leading to the solution  $c \equiv 1027 \pmod{2856}$ , which, together with the additional information, implies that there were exactly 3883 gymnasts.  $\square$

**11.D.6.** Find which of the following (systems of) linear congruences has a solution.

- i)  $x \equiv 1 \pmod{3}$ ,  
 $x \equiv -1 \pmod{9}$ ;
- ii)  $8x \equiv 1 \pmod{12345678910111213}$ ;
- iii)  $x \equiv 3 \pmod{29}$ ,  
 $x \equiv 5 \pmod{47}$ .  $\circ$

The Chinese remainder theorem can also be used “in the opposite direction”, i.e. to simplify a linear congruence provided we are able to express the modulus as a product of pairwise coprime factors.

**11.D.7.** Solve the congruence  $23\,941x \equiv 915 \pmod{3564}$ .

**Solution.** Let us factor  $3564 = 2^2 \cdot 3^4 \cdot 11$ . Since none of the integers 2, 3, 11 divides 23 941, we have  $(23\,941, 3564) = 1$ , so the congruence has a solution. Since  $\varphi(3564) = 2 \cdot (3^3 \cdot 2) \cdot 10 = 1080$ , the solution is of the form  $x \equiv 915 \cdot 23\,941^{1079} \pmod{3564}$ . However, it would take much effort to simplify the right-hand integer to a more explicit form. Therefore, we will try to solve the congruence in a different way – we will build an equivalent system of congruences which are easier to solve than the original one.

We know that an integer  $x$  is a solution of the given congruence if and only if it is a solution of the system

$$\begin{aligned} 23941x &\equiv 915 \pmod{2^2}, \\ 23941x &\equiv 915 \pmod{3^4}, \\ 23941x &\equiv 915 \pmod{11}. \end{aligned}$$

Solving these congruences separately, we get the following, equivalent system:

$$\begin{aligned} x &\equiv 3 \pmod{4}, \\ x &\equiv -3 \pmod{81}, \\ x &\equiv -4 \pmod{11}. \end{aligned}$$

Now, the procedure for finding a solution of a system of congruences yields  $x \equiv -1137 \pmod{3564}$ , which is the solution of the original congruence as well.  $\square$

**11.D.8.** Solve the congruence  $3446x \equiv 8642 \pmod{208}$ .  $\circ$

**11.D.9.** Prove that the sequence  $(2^n - 3)_{n=1}^\infty$  contains infinitely many multiples of 5 as well as infinitely many multiples of 13, yet there is no multiple of 65 in it.  $\circ$

**Residue number system.** When calculating with large integers, it is often more advantageous to work not with their decimal or binary expansions, but rather with their representation in a so-called *residue number system*, which allows for easy parallelization of computations with large integers. Such a system is given by a  $k$ -tuple of (usually pairwise coprime) moduli, and each integer which



At the same time (utilizing the previous lemma), the order of every element of  $S$  apparently divides  $2^{\ell-2}$ . Therefore, this reduced system cannot (and nor can any other) contain an element of order  $\varphi(2^\ell) = 2^{\ell-1}$ .  $\square$

**11.3.19.** The last piece to the jigsaw puzzle of propositions which collectively prove theorem 11.3.16 is the statement about non-existence of primitive roots for composite numbers which are neither a power of prime nor twice such.

**Proposition.** Let  $m \in \mathbb{N}$  be divisible by at least two primes, and let it not be twice a power of an odd prime. Then, there are no primitive roots modulo  $m$ .

**PROOF.** Let  $m$  factor to primes as  $2^\alpha p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ , where  $\alpha \in \mathbb{N}_0, \alpha_i \in \mathbb{N}, 2 \nmid p_i$ , and  $k \geq 2$  or both  $k \geq 1$  and  $\alpha \geq 2$ . Denoting  $\delta = [\varphi(2^\alpha), \varphi(p_1^{\alpha_1}), \dots, \varphi(p_1^{\alpha_1})]$ , we can easily see that  $\delta < \varphi(2^\alpha) \cdot \varphi(p_1^{\alpha_1}) \cdots \varphi(p_1^{\alpha_1}) = \varphi(m)$  and that for any  $a \in \mathbb{Z}, (a, m) = 1$ , we have  $a^\delta \equiv 1 \pmod{m}$ . Therefore, there are no primitive roots modulo  $m$ .  $\square$

**11.3.20.** In general, it is computationally very hard to find a primitive root for a given modulus. The following theorem describes a necessary and sufficient condition for the examined integer to be a primitive root.



**Theorem.** Let  $m$  be such an integer that there are primitive roots modulo  $m$ . Let us write  $\varphi(m) = q_1^{\alpha_1} \cdots q_k^{\alpha_k}$ , where  $q_1, \dots, q_k$  are primes and  $\alpha_1, \dots, \alpha_k \in \mathbb{N}$ . Then, for every  $g \in \mathbb{Z}, (g, m) = 1$ , it holds that  $g$  is a primitive root modulo  $m$ , if and only if neither of the following congruences holds:

$$g^{\frac{\varphi(m)}{q_1}} \equiv 1 \pmod{m}, \dots, g^{\frac{\varphi(m)}{q_k}} \equiv 1 \pmod{m}.$$

**PROOF.** If either of the congruences were true, it would mean that the order of  $g$  is less than  $\varphi(m)$ .

On the other hand, if  $g$  fails to be a primitive root, then there is a  $d \in \mathbb{N}, d \mid \varphi(m)$ , where  $d < \varphi(m)$  and  $g^d \equiv 1 \pmod{m}$ . If  $u = \frac{\varphi(m)}{d} > 1$ , then there must be an  $i \in \{1, \dots, k\}$  such that  $q_i \mid u$ . However, then we get

$$g^{\frac{\varphi(m)}{q_i}} = g^{d \cdot \frac{u}{q_i}} \equiv 1 \pmod{m}.$$

$\square$

#### 4. Solving congruences and systems of them

This part will be devoted to the analog to solving equations in a numerical domain. We will actually be solving equations (and systems of equations) in the ring of residue classes  $(\mathbb{Z}_m, +, \cdot)$ ; we will, however, talk about solving congruences modulo  $m$  and write it in the more transparent way as usual.



is less than their product is then uniquely representable as a  $k$ -tuple of remainders (whose values do not exceed the moduli).

**11.D.10.** The quintuple of moduli 3, 5, 7, 11, 13 can serve to uniquely represent integers which are less than their product (i.e. less than 15015) and to perform standard arithmetic operations efficiently (and in a distributed manner if desired). Now, we will determine the representation of the integers 1234 and 5678 in this residue number system and we will determine their sum and product.

**Solution.** Calculating the remainders of the given integers upon division by the particular moduli, we get their RNS representations, which can be written as the tuples (1, 4, 2, 2, 12) and (2, 3, 1, 2, 10).

The sum is computed componentwise (reducing the results modulo the appropriate number), leading to the tuple (0, 2, 3, 4, 9). Using the Chinese remainder theorem, this tuple can then be transformed back to the integer 6912. The product is computed analogously, yielding the corresponding tuple (2, 2, 2, 4, 3), which can be transformed back to 9662 (by the Chinese remainder theorem again). This is indeed congruent to  $1234 \cdot 5678$  modulo 15015.  $\square$

**11.D.11.** In practice, the residue number system is often a triple  $2^n - 1, 2^n, 2^n + 1$  (why are these integers always coprime?), which can uniquely cover integers of  $3n$  bits at the utmost.

Consider the case  $n = 3$  and determine the representation of the integer 118 in this residue number system.

**Solution.** We can directly calculate that  $118 \equiv 6 \pmod{7}$ ,  $118 \equiv 6 \pmod{8}$ , and  $118 \equiv 1 \pmod{9}$ . The wanted representation is thus given by the triple (6, 6, 1).

In practice, however, it is very important that the RNS representation can be efficiently transformed to binary and vice versa. In our concrete case, the remainder of  $118 = (1110110)_2$  when divided by  $2^3$  can be found easily – it is the last three bits  $(110)_2 = 6$ . Computing the remainder upon division by  $2^3 + 1 = 9$  or  $2^3 - 1 = 7$  is not any more complicated. We can see (splitting the examined integer into three groups of  $n$  bits each) that

$$(1110110)_2 \equiv (001)_2 + (110)_2 + (110)_2 \equiv 6 \pmod{2^3 - 1},$$

$$(1110110)_2 \equiv (001)_2 - (110)_2 + (110)_2 \equiv 1 \pmod{2^3 + 1}.$$

A thoughtful reader has surely noticed the similarity with the criteria for divisibility by 9 and 11, which were discussed in paragraph 11.C.9.  $\square$

**11.D.12. Higher-order congruences.** Using the procedure of theorem 11.4.6, solve the congruence



$$x^4 + 7x + 4 \equiv 0 \pmod{27}.$$

**Solution.** First, we will solve this congruence modulo 3 (by substitution, for instance) – we can easily find that the solution

CONGRUENCE IN ONE VARIABLE

Let  $m \in \mathbb{N}$ ,  $f(x), g(x) \in \mathbb{Z}[x]$ . The notation

$$f(x) \equiv g(x) \pmod{m}$$

is called a *congruence in variable  $x$* , and it is understood to be the problem of finding the *set of solutions*, i.e., the set of all such integers  $c$  for which  $f(c) \equiv g(c) \pmod{m}$ .

Two congruences (in one variable) are called *equivalent* if and only if they have the same set of solutions.

The mentioned congruence is equivalent to the congruence

$$f(x) - g(x) \equiv 0 \pmod{m}.$$

The only method which always leads to a solution is trying out all possible values (however, this would, of course, often take too much time). This procedure is formalized by the following proposition.

**11.4.1. Proposition.** Let  $m \in \mathbb{N}$ ,  $f(x) \in \mathbb{Z}[x]$ . Then, it holds for every  $a, b \in \mathbb{Z}$  that

$$a \equiv b \pmod{m} \implies f(a) \equiv f(b) \pmod{m}.$$

**PROOF.** Let  $f(x) = c_n x^n + c_{n-1} x^{n-1} + \dots + c_1 x + c_0$ , where  $c_0, c_1, \dots, c_n \in \mathbb{Z}$ . Since  $a \equiv b \pmod{m}$ ,  $c_i a^i \equiv c_i b^i \pmod{m}$  holds for every  $i = 0, 1, \dots, n$ . Adding up these congruences for  $i = 0, 1, 2, \dots, n$  leads to

$$c_n a^n + \dots + c_1 a + c_0 \equiv c_n b^n + \dots + c_1 b + c_0 \pmod{m},$$

i.e.,  $f(a) \equiv f(b) \pmod{m}$ .  $\square$

**Corollary.** The set of solutions of an arbitrary congruence modulo  $m$  is a union of residue classes modulo  $m$ .

**Definition.** The number of solutions of a congruence in one variable modulo  $m$  is the number of residue classes modulo  $m$  containing the solutions of the congruence.

**Example.** The concept *number of solutions of a congruence*, which we have just defined, is a bit counterintuitive in that it depends on the modulus of the congruence. Therefore, equivalent congruences (sharing the same integers as solutions) can have different numbers of solutions.

- (1) The congruence  $2x \equiv 3 \pmod{3}$  has exactly one solution (modulo 3).
- (2) The congruence  $10x \equiv 15 \pmod{15}$  has five solutions (modulo 15).
- (3) The congruences from (1) and (2) are equivalent.

**11.4.2. Linear congruence in one variable.** Just like in the case of ordinary equations, the easiest congruences are the linear ones, for which we are able not only to decide whether they have a solution, but to efficiently find it (provided they have some). The procedure is described by the following theorem and its proof.

**11.4.3. Theorem.** Let  $m \in \mathbb{N}$ ,  $a, b \in \mathbb{Z}$ , and  $d = (a, m)$ . Then the congruence (in variable  $x$ )

$$ax \equiv b \pmod{m}$$



is  $x \equiv 1 \pmod{3}$ . Now, writing the solution in the form  $x = 1 + 3t$ , where  $t \in \mathbb{Z}$ , we will solve the congruence modulo 9:

$$\begin{aligned} x^4 + 7x + 4 &\equiv 0 \pmod{9}, \\ (1 + 3t)^4 + 7(1 + 3t) + 4 &\equiv 0 \pmod{9}, \\ 1 + 4 \cdot 3t + 7 + 7 \cdot 3t + 4 &\equiv 0 \pmod{9}, \\ 33t &\equiv -12 \pmod{9}, \\ 11t &\equiv -4 \pmod{3}, \\ t &\equiv 1 \pmod{3}. \end{aligned}$$

Writing  $t = 1 + 3s$ , where  $s \in \mathbb{Z}$ , we get  $x = 4 + 9s$ , and substituting this leads to

$$\begin{aligned} (4 + 9s)^4 + 7(4 + 9s) + 4 &\equiv 0 \pmod{27}, \\ 4^4 + 4 \cdot 4^3 \cdot 9s + 28 + 63s + 4 &\equiv 0 \pmod{27}, \\ 256 \cdot 9s + 63s &\equiv -288 \pmod{27}, \\ 256s + 7s &\equiv -32 \pmod{3}, \\ 2s &\equiv 1 \pmod{3}, \\ s &\equiv 2 \pmod{3}. \end{aligned}$$

Altogether, we get the solution in the form  $x = 4 + 9s = 4 + 9(2 + 3r) = 22 + 27r$ , where  $r \in \mathbb{Z}$ , i.e.,  $x \equiv 22 \pmod{27}$ .  $\square$

**11.D.13.** Knowing a primitive root modulo 41 from exercise 11.C.29, solve the congruence

$$7x^{17} \equiv 11 \pmod{41}.$$

**Solution.** Multiplying the congruence by 6, we get an equivalent congruence  $42x^{17} \equiv 66$ , i.e.,  $x^{17} \equiv 25 \pmod{41}$ . Since 6 is a primitive root modulo 41, the substitution  $x = 6^t$  leads to the congruence  $6^{17t} \equiv 25 \equiv 6^4 \pmod{41}$ , which is equivalent to  $17t \equiv 4 \pmod{40}$ , and this holds if and only if  $t \equiv 12 \pmod{40}$ . Therefore, the congruence is satisfied by exactly those integers  $x$  with  $x \equiv 6^{12} \equiv 4 \pmod{41}$ .  $\square$

**11.D.14.** Solve the congruence  $x^5 + 1 \equiv 0 \pmod{11}$ .

**Solution.** Since  $(5, \varphi(11)) = 5$  and

$$(-1)^{\frac{\varphi(11)}{5}} \equiv 1 \pmod{11},$$

the congruence

$$x^5 \equiv -1 \pmod{11}$$

has five solutions. There are several possibilities how to find them. We can either try all (ten) candidates or transform the problem to a linear congruence using the primitive-root trick. Since  $2^{10/2} \equiv -1 \not\equiv 1 \pmod{11}$  and  $2^{10/5} \equiv 4 \not\equiv 1 \pmod{11}$ , 2 is a primitive root modulo 11 (see also exercise 11.C.28), and the substitution  $x \equiv 2^y$  then transforms the congruence to

$$2^{5y} \equiv 2^5 \pmod{11},$$

has a solution if and only if  $d \mid b$ .

If  $d \mid b$ , then this congruence has exactly  $d$  solutions (modulo  $m$ ).

**PROOF.** First, we prove that the mentioned condition is necessary. If an integer  $c$  is a solution of this congruence, then we must have  $m \mid a \cdot c - b$ . Since  $d = (a, m)$ , we get  $d \mid m$  and  $d \mid a \cdot c - b$ , so  $d \mid a \cdot c - (a \cdot c - b) = b$ .

Now, we will prove that if  $d \mid b$ , then the given congruence has exactly  $d$  solutions modulo  $m$ . Let  $a_1, b_1 \in \mathbb{Z}$  and  $m_1 \in \mathbb{N}$  so that  $a = d \cdot a_1$ ,  $b = d \cdot b_1$ , and  $m = d \cdot m_1$ . The congruence we are trying to solve is thus equivalent to the congruence

$$a_1 \cdot x \equiv b_1 \pmod{m_1},$$

where  $(a_1, m_1) = 1$ . This congruence can be multiplied by the integer  $a_1^{\varphi(m_1)-1}$ , which, by Euler's theorem, leads to

$$x \equiv b_1 \cdot a_1^{\varphi(m_1)-1} \pmod{m_1}.$$

This congruence has a unique solution modulo  $m_1$ , thus it has  $d = m/m_1$  solutions modulo  $m$ .  $\square$

Using the theorem about solutions of linear congruences, we can, among others, prove Wilson's theorem – an important theorem which gives a necessary (and sufficient) condition for an integer to be a prime. Such conditions are extremely useful in computational number theory, where one needs to efficiently determine whether a given large integer is a prime. Unfortunately, it is not known now how fast modular factorial of a large integer can be computed. That is why Wilson's theorem is not used for this purpose in practice.

**Theorem (Wilson).** A natural number  $n > 1$  is a prime if and only if

$$(n - 1)! \equiv -1 \pmod{n}.$$

**PROOF.** First, we prove that every composite number  $n > 4$  satisfies  $n \mid (n - 1)!$ , i.e.,  $(n - 1)! \equiv 0 \pmod{n}$ . Let  $1 < d < n$  be a non-trivial divisor of  $n$ . If  $d \neq n/d$ , then the inequality  $1 < d, n/d \leq n - 1$  implies what we need:  $n = d \cdot n/d \mid (n - 1)!$ . If  $d = n/d$ , i.e.,  $n = d^2$ , then we have  $d > 2$  (since  $n > 4$ ) and  $n \mid (d \cdot 2d) \mid (n - 1)!$ . For  $n = 4$ , we easily get  $(4 - 1)! \equiv 2 \not\equiv -1 \pmod{4}$ .

Now, let  $p$  be a prime. The integers in the set  $\{2, 3, \dots, p - 2\}$  can be grouped by pairs of those mutually inverse modulo  $p$ , i.e., pairs of integers whose product is congruent to 1. By the previous theorem, for every integer  $a$  of this set, there is a unique solution of the congruence  $a \cdot x \equiv 1 \pmod{p}$ . Since  $a \neq 0, 1, p - 1$ , it is apparent that the solution  $c$  of the congruence also satisfies  $c \not\equiv 0, 1, -1 \pmod{p}$ . The integer  $a$  cannot be paired with itself, either: If so, i.e.,  $a \cdot a \equiv 1 \pmod{p}$ , we would (thanks to  $p \mid a^2 - 1 = (a + 1)(a - 1)$ ) get the congruence  $a \equiv \pm 1 \pmod{p}$ . The product of the integers of the mentioned set thus consists of products of  $(p - 3)/2$  pairs (whose product is always congruent to 1 modulo  $p$ ). Therefore, we have

$$(p - 1)! \equiv 1^{(p-3)/2} \cdot (p - 1) \equiv -1 \pmod{p}. \quad \square$$

which is equivalent to the linear congruence

$$\begin{aligned} 5y &\equiv 5 \pmod{10}, \\ y &\equiv 1 \pmod{2}. \end{aligned}$$

This congruence is satisfied by  $y \in \{-3, -1, 1, 3, 5\}$ ; the original congruence is thus (substituting  $x \equiv 2^y \pmod{11}$ ) satisfied by  $x \in \{-1, 2, -3, -4, -5\}$ .  $\square$

**11.D.15.** Solve the congruence  $x^3 - 3x + 5 \equiv 0 \pmod{105}$ .  $\circ$

**11.D.16.** Determine the number of solutions of the congruence



$$x^5 \equiv 534 \pmod{23^2}.$$

**Solution.** The given congruence is equivalent to  $x^5 \equiv 5 \pmod{23^2}$ , and since we have  $(5, \varphi(23)) = 1$ , it follows from the theorem on solvability of binomial congruences that the congruence has a unique solution if considered modulo 23. Furthermore, this solution is surely not a multiple of 23. Therefore, considering the polynomial whose roots we are looking for, its derivative  $(x^5 - 5)' = 5x^4$  does not evaluate to a multiple of 23 at the wanted solution, either. Invoking Hensel's lemma, we can summarize that the original congruence has a unique solution (without having to describe it explicitly).  $\square$

**11.D.17.** Give an example of a polynomial congruence whose degree is less than the number of its solutions.

**Solution.** Taking into account theorem 11.4.8, we must use either a modulus which is composite or a polynomial all of whose coefficients will be multiples of the modulus.

As an example of a congruence of the first kind, we can put

$$x^2 \equiv 1 \pmod{8},$$

which is a quadratic congruence with four solutions 1, 3, 5, 7.

The case if a prime modulus can be exemplified by the quadratic congruence  $10x^2 - 15 \equiv 0 \pmod{5}$ , which has five solutions.  $\square$

**11.D.18. Other types of congruences.** Prove that for any natural number  $n$ , the integer



$$111 + 2^{2^{2n-1}}$$

is divisible by 127.

**Solution.** We are to prove that the congruence

$$2^{2^{2n-1}} \equiv -111 \pmod{127}$$

is satisfied for every  $n \in \mathbb{N}$ . This congruence is equivalent to

$$2^{2^{2n-1}} \equiv 2^2 \pmod{127}.$$

Since  $2^7 = 128 \equiv 1 \pmod{127}$ , the order of 2 modulo 127 equals 7, so the congruence to be proved is (by 11.3.13) equivalent to

$$2^{2^{2n-1}} \equiv 2^2 \pmod{7}.$$

**11.4.4. Systems of linear congruences.** Having a system of linear congruences in the same variable, we can decide whether each of them is solvable by the previous theorem. If at least one of the congruences does not have a solution, nor does the whole system. On the other hand, if each of the congruences is solvable, we can rearrange it into the form  $x \equiv c_i \pmod{m_i}$ .



We thus get a system of congruences

$$\begin{aligned} x &\equiv c_1 \pmod{m_1}, \\ &\vdots \\ x &\equiv c_k \pmod{m_k}. \end{aligned}$$

Apparently, it suffices to solve the case  $k = 2$  since the solutions of a system of more congruences can be obtained by repeatedly applying the procedure for a system of two congruences.

**Proposition.** Let  $c_1, c_2$  be integers and  $m_1, m_2$  be natural numbers. Let us denote  $d = (m_1, m_2)$ . The system of two congruences

$$\begin{aligned} x &\equiv c_1 \pmod{m_1}, \\ x &\equiv c_2 \pmod{m_2} \end{aligned}$$

has no solution if  $c_1 \not\equiv c_2 \pmod{d}$ . On the other hand, if  $c_1 \equiv c_2 \pmod{d}$ , then there is an integer  $c$  such that  $x \in \mathbb{Z}$  satisfies the system if and only if it satisfies the congruence

$$x \equiv c \pmod{[m_1, m_2]}.$$

**PROOF.** If the given system is to have a solution  $x \in \mathbb{Z}$ , we must have  $x \equiv c_1 \pmod{d}$ ,  $x \equiv c_2 \pmod{d}$ , and thus  $c_1 \equiv c_2 \pmod{d}$  as well. Hence it follows that the system cannot have a solution when  $c_1 \not\equiv c_2 \pmod{d}$ .

From now on, suppose that  $c_1 \equiv c_2 \pmod{d}$ . The first congruence of the system is satisfied by those integers  $x$  which are of the form  $x = c_1 + tm_1$ , where  $t \in \mathbb{Z}$  is arbitrary. Such an integer  $x$  satisfies the second congruence of the system if and only if  $c_1 + tm_1 \equiv c_2 \pmod{m_2}$ , i.e.,  $tm_1 \equiv c_2 - c_1 \pmod{m_2}$ . By the theorem about solutions of linear congruences, this congruence (in variable  $t$ ) is solvable since  $d = (m_1, m_2)$  divides  $c_2 - c_1$ , and  $t$  satisfies this congruence if and only if

$$t \equiv \frac{c_2 - c_1}{d} \cdot \left(\frac{m_1}{d}\right)^{\varphi\left(\frac{m_2}{d}\right)-1} \pmod{\frac{m_2}{d}},$$

i.e., if and only if

$$\begin{aligned} x &= c_1 + tm_1 = c_1 + (c_2 - c_1) \cdot \left(\frac{m_1}{d}\right)^{\varphi\left(\frac{m_2}{d}\right)} + r \frac{m_1 m_2}{d} \\ &= c + r \cdot [m_1, m_2], \text{ where } r \in \mathbb{Z} \text{ is arbitrary and} \end{aligned}$$

$c = c_1 + (c_2 - c_1) \cdot (m_1/d)^{\varphi(m_2/d)}$ , as  $m_1 m_2$  equals  $d \cdot [m_1, m_2]$ . We have thus found such an integer  $c$  that every  $x \in \mathbb{Z}$  satisfies the system if and only if  $x \equiv c \pmod{[m_1, m_2]}$ , as wanted.  $\square$



Similarly, the order of 2 modulo 7 is 3, which leads to the (again equivalent) congruence

$$\begin{aligned} 2^{2n-1} &\equiv 2 \pmod{3}, \\ (-1)^{2n-1} &\equiv -1 \pmod{3}, \end{aligned}$$

and this is apparently true (we could also have proceed likewise – the order of 2 modulo 3 is 2, and so on). This proves the statement.  $\square$

**11.D.19.** Determine which natural numbers  $n$  satisfy that the integer  $n \cdot 2^n + 1$  is divisible by seven.



**Solution.** We are looking for the solution of the congruence

$$n \cdot 2^n \equiv -1 \pmod{7}.$$

We should be aware of the fact that we cannot use the theorem 11.4.1 since  $n \cdot 2^n$  is not a polynomial in variable  $n$ , so it is not guaranteed (and it is even not true) that the expression will yield the same remainder modulo 7 when evaluated at integers which are congruent modulo 7.

On the other hand, we can notice that the order of 2 modulo 7 is equal to 3, so we can split the problem into three cases according to the remainder of  $n$  when divided by 3.

For  $n \equiv 0 \pmod{3}$ , we have  $2^n \equiv 1 \pmod{7}$ , so the congruence in question is equivalent to  $n \equiv -1 \pmod{7}$ . Combining the conditions  $n \equiv 0 \pmod{3}$  and  $n \equiv -1 \pmod{7}$  in the Chinese remainder theorem leads to the solution  $n \equiv 6 \pmod{21}$ .

Now, for  $n \equiv 1 \pmod{3}$ , we have  $2^n \equiv 2 \pmod{7}$ , so the examined congruence is of the form  $2n \equiv -1 \pmod{7}$ , which is equivalent to  $n \equiv 3 \pmod{7}$ . The conditions  $n \equiv 1 \pmod{3}$  and  $n \equiv 3 \pmod{7}$  are satisfied iff  $n \equiv 10 \pmod{21}$ .

Finally, for  $n \equiv 2 \pmod{3}$ , we have  $2^n \equiv 4 \pmod{7}$ , and the solution of the congruence  $4n \equiv -1 \pmod{7}$  is  $n \equiv 5 \pmod{7}$ . Altogether,  $n \equiv 5 \pmod{21}$ .

The problem is satisfied by exactly those natural numbers  $n$  with  $n \equiv 5, 6, 10 \pmod{21}$ .  $\square$

**11.D.20.** Prove that for any natural number  $n$ , the integer  $2n^4 + n^3 + 50$  is divisible by 6 if and only if the integer  $2 \cdot 4^n + 3^n + 50$  is divisible by 13.



**Solution.** The expression  $f(n) = 2n^4 + n^3 + 50$  is a polynomial in variable  $n$ , so in this case, we can make use of theorem 11.4.1, i.e., it suffices to go through all possible remainders modulo 6. Since the order of 4 modulo 13 is equal to 6 and the order of 3 modulo 13 equals 3, it is enough (by 11.3.13) to examine the remainder of  $n$  upon division by 6 in the latter case as well.

In the former case, we calculate

$n$	0	1	2	3	4	5
$f(n) \pmod{6}$	2	5	0	5	2	3

We can notice that the proof of this theorem is constructive, i.e., it yields a formula for finding the integer  $c$ . This theorem thus gives us a procedure how to catch the condition that an integer  $x$  satisfies a given system by a single congruence. This new congruence is then of the same form as the original one. Therefore, we can apply this procedure to a system of more congruences – first, we create a single congruence from the first and second congruences of the system (satisfied by exactly those integers  $x$  which satisfy the original two); then, we create another congruence from the new one and the third one of the original system, and so on. Each step reduces the number of congruences by one; after a finite number of steps, we thus arrive at a single congruence which describes all solutions of the given system.

It follows from the procedure we have just mentioned (supposing the condition from below holds) that a system of congruences always has a solution, and this is unique.

**Theorem (Chinese remainder theorem).** Let  $m_1, \dots, m_k \in \mathbb{N}$  be pairwise coprime,  $a_1, \dots, a_k \in \mathbb{Z}$ . Then, the system

$$\begin{aligned} x &\equiv a_1 \pmod{m_1}, \\ &\vdots \\ x &\equiv a_k \pmod{m_k} \end{aligned}$$

has a unique solution modulo  $m_1 \cdot m_2 \cdot \dots \cdot m_k$ .

**Remark.** The unusual name of this theorem comes from Chinese mathematician Sun Tzu of the 4th century. In his text, he asked for an integer which leaves remainder 2 when divided by 3, leaves remainder 3 when divided by 5, and again remainder 2 when divided by 7.

The answer is rumored to be hidden in the following song:

孫子歌 Sunzi Ge

三人同行七十里  
五樹梅花廿一枝  
七子團圓正月半  
一百零五轉回起

**PROOF.** It is a simple consequence of the previous proposition about the form of the solution of a system of two congruences. However, as we show here, this result can also be proved directly. Let us denote  $M := m_1 m_2 \cdot \dots \cdot m_r$  and  $n_i = M/m_i$  for every  $i$ ,  $1 \leq i \leq r$ . Then, for any  $i$ ,  $m_i$  is coprime to  $n_i$ , so there is an integer  $b_i \in \{1, \dots, m_i - 1\}$  such that  $b_i n_i \equiv 1 \pmod{m_i}$ . Note that  $b_i n_i$  is divisible by all the numbers  $m_j$ ,  $1 \leq j \leq r$ ,  $i \neq j$ . Therefore, the wanted solution of the system is the integer

$$x = a_1 b_1 n_1 + a_2 b_2 n_2 + \dots + a_r b_r n_r.$$

$\square$

Therefore, the congruence  $f(n) \equiv 0 \pmod{6}$  is satisfied by exactly those natural numbers  $n$  which satisfy  $n \equiv 2 \pmod{6}$ .

In the latter case, we gradually compute that

$n$	0	1	2	3	4	5
$4^n \pmod{13}$	1	4	3	-1	-4	-3
$3^n \pmod{13}$	1	3	9	1	3	9
$2 \cdot 4^n + 3^n - 2 \pmod{13}$	1	9	0	-3	-7	1

Just like in the former case, the congruence  $2 \cdot 4^n + 3^n + 50 \equiv 0 \pmod{13}$  is satisfied if and only if  $n \equiv 2 \pmod{6}$ .  $\square$

**11.D.21.** Solve the congruence

$$x^2 \equiv 18 \pmod{63}.$$

**Solution.** Since  $(18, 63) = 9$ , it must be that  $9 \mid x^2$ , i.e.,  $3 \mid x$ . Setting  $x = 3x_1$ ,  $x_1 \in \mathbb{Z}$ , we get an equivalent congruence  $x_1^2 \equiv 2 \pmod{7}$ , which already satisfies that the modulus is coprime to the integer on the right-hand side. It follows from theorem 11.4.8 that this congruence has at most 2 solutions, and those are clearly  $x_1 \equiv \pm 3 \pmod{7}$ , i.e.,  $x_1 \equiv \pm 3, \pm 10, \pm 17, \pm 24, \pm 31, \pm 38, \pm 45, \pm 52, \pm 59 \pmod{63}$ . The solution of the original congruence is thus  $x \equiv 3x_1 \pmod{63}$ , i.e.,  $x \equiv \pm 9, \pm 12, \pm 30 \pmod{63}$ .  $\square$

**11.D.22.** Solve the congruence

$$x^3 \equiv 3 \pmod{18}.$$

**Solution.** Since  $(3, 18) = 3$ , we must have  $3 \mid x$ . Making the substitution  $x = 3 \cdot x_1$ , similarly to the above exercise, we get the congruence

$$27x_1^3 \equiv 3 \pmod{18},$$

which has no solution since  $(27, 18) \nmid 3$ .  $\square$

**Quadratic congruences.** In the theoretical column we state that any quadratic congruence can be transformed to the (possibly system of congruences of) binomial form  $x^2 \equiv a \pmod{p}$ , and then we can decide about the solvability using the Legendre symbol. Let us illustrate it on several examples.



**11.D.23.** Determine the number of solutions of the congruence  $13x^2 + 7x + 1 \equiv 0 \pmod{37}$ .

**Solution.** First, we need to normalize the polynomial on the left-hand side, i.e. we have to find the inverse of 13 modulo 37. Using the Euclidean algorithm we find that the inverse is 20, and after multiplication of both sides of the congruence by it and reducing modulo 37 we obtain the congruence  $x^2 + 29x + 20 \equiv 0 \pmod{37}$ . Now we complete the square (an odd coefficient 29 does not cause any trouble as it can be replaced by  $-8$ ) and we obtain  $(x - 4)^2 + 4 \equiv 0 \pmod{37}$ . After substitution of  $y$  for  $x - 4$  we finally obtain the congruence in a binomial form

$$y^2 \equiv -4 \pmod{37}.$$

Let us emphasize that this is quite a strong theorem (which is actually valid in much more general algebraic structures), which allows us to guarantee that for any remainders with respect to given (pairwise coprime) moduli, there exists an integer with the given remainders.

**11.4.5. Higher-order congruences.** Now, let us get back to the more general case of congruences.



$$f(x) \equiv 0 \pmod{m},$$

where  $f(x)$  is a polynomial with integer coefficients and  $m \in \mathbb{N}$ . So far, we have only one method at our disposal, which is tedious, yet universal – to try all possible remainders modulo  $m$ . When solving such a congruence, it is sufficient to find out for which integers  $a$ ,  $0 \leq a < m$ , it holds that  $f(a) \equiv 0 \pmod{m}$ . The disadvantage of this method is its complexity, which increases as  $m$  does. If  $m$  is composite, i.e.,  $m = p_1^{\alpha_1} \dots p_k^{\alpha_k}$ , where  $p_1, \dots, p_k$  are distinct primes, and  $k > 1$ , we can replace the original congruence by the system of congruences

$$f(x) \equiv 0 \pmod{p_1^{\alpha_1}},$$

$\vdots$

$$f(x) \equiv 0 \pmod{p_k^{\alpha_k}},$$

which has the same set of solutions. However, we can solve the congruences separately. The advantage of this method is in that the moduli of the congruences of the system are less than the modulus of the original congruence.

**Example.** Consider the congruence

$$x^3 - 2x + 11 \equiv 0 \pmod{105}.$$

If we were to try out all possibilities, we would have to compute the value of  $f(x) = x^3 - 2x + 11$  for the 105 values  $f(0), f(1), \dots, f(104)$ . Therefore, we better factor  $105 = 3 \cdot 5 \cdot 7$  and solve the congruences  $f(x) \equiv 0$  for moduli 3, 5, and 7. We evaluate the polynomial  $f(x)$  in convenient integers:

$x$	-3	-2	-1	0	1	2	3
$f(x)$	-10	7	12	11	10	15	32

The congruence  $f(x) \equiv 0 \pmod{3}$  thus has solution  $x \equiv -1 \pmod{3}$  (only the first one of the integers 12, 11, 10 is a multiple of 3); the congruence  $f(x) \equiv 0 \pmod{5}$  has solutions  $x \equiv 1$  and  $x \equiv 2 \pmod{5}$ ; finally, the solution of the congruence  $f(x) \equiv 0 \pmod{7}$  is  $x \equiv -2 \pmod{7}$ .

It remains to solve two systems of congruences:

$$\begin{array}{ll} x \equiv -1 \pmod{3}, & x \equiv -1 \pmod{3}, \\ x \equiv 1 \pmod{5}, & \text{and } x \equiv 2 \pmod{5}, \\ x \equiv -2 \pmod{7} & x \equiv -2 \pmod{7}. \end{array}$$

Solving these systems, we can find out that the solutions of the given congruence  $f(x) \equiv 0 \pmod{105}$  are exactly those integers  $x$  which satisfy  $x \equiv 26 \pmod{105}$  or  $x \equiv 47 \pmod{105}$ .

The fact that this congruence is solvable can be established either using theorem 11.4.10, or with use of the Legendre symbol. The former approach leads to the calculation  $d = (2, \varphi(37)) = 2$ , and

$$(-4)^{\frac{36}{2}} \equiv 1 \pmod{37},$$

while the latter one gives

$$\left(\frac{-4}{37}\right) = \left(\frac{-1}{37}\right) \cdot \left(\frac{2}{37}\right)^2 = 1$$

by the corollary after theorem 11.4.13 (as  $37 \equiv 1 \pmod{4}$ ). In any way we have obtained that the given congruence has  $d = 2$  solutions.  $\square$

**11.D.24.** Solve the congruence  $6x^2 + x - 1 \equiv 0 \pmod{29}$ .

**Solution.** Although we have not presented any special method for finding solutions of quadratic congruence yet (apart from the general method for binomial congruences or going through the complete residue system) we will see that in some case the set of solutions can be easily established. Let us first proceed in the usual way: multiplying the congruence by 5 (it is the inverse of 6 modulo 29) we obtain  $x^2 + 5x - 5 \equiv 0 \pmod{29}$ , and after completing the square we have

$$(x - 12)^2 \equiv 4 \pmod{29}.$$

We immediately see that this congruence is solvable with the pair of solutions  $x - 12 \equiv \pm 2 \pmod{29}$ , and thus  $x \equiv 10, 14 \pmod{29}$ .

We could have also seen almost immediately that the given polynomial can be factored as  $6x^2 + x - 1 = (3x - 1)(2x + 1)$ , and thus the prime modulus 29 has to divide either  $3x - 1$  or  $2x + 1$ . The obtained linear congruences  $3x \equiv 1 \pmod{29}$  and  $2x \equiv -1 \pmod{29}$  easily yield the same solutions  $x \equiv 10 \pmod{29}$  and  $x \equiv 14 \pmod{29}$  as above.  $\square$

**11.D.25.** Find all integers which satisfy the congruence

$$x^2 \equiv 7 \pmod{43}.$$

**Solution.** The Legendre symbol evaluates to

$$\left(\frac{7}{43}\right) = -\left(\frac{43}{7}\right) = -\left(\frac{1}{7}\right) = -1.$$

Hence it follows that 7 is a quadratic nonresidue modulo 43, so there is no solution of the given congruence.  $\square$

**11.D.26.** Find all integers  $a$  for which the congruence

$$x^2 \equiv a \pmod{43}$$

is solvable.

**Solution.** This exercise is a follow-up to the above one, from which we can see that the integer 7 does not meet the requirement. We can test all the remainders modulo 43 in the same way, but there is a simpler method. The congruence is surely solvable if  $a$  is a multiple of 43 (then, it has a unique solution);

It is not always possible to replace the congruence with a system of congruences modulo primes, as in the above example: if the original modulus is a multiple of a higher power of a prime, then we cannot “get rid” of this power. However, even such a congruence modulo a power of prime need not be solved by examining all possibilities. There is a more efficient tool, which is described by the following theorem.



**11.4.6. Theorem (Hensel’s lemma).** *Let  $p$  be a prime,  $f(x) \in \mathbb{Z}[x]$ ,  $a \in \mathbb{Z}$  such that  $p \mid f(a), p \nmid f'(a)$ . Then, for every  $n \in \mathbb{N}$ , the system*

$$\begin{aligned} x &\equiv a \pmod{p}, \\ f(x) &\equiv 0 \pmod{p^n} \end{aligned}$$

*has a unique solution modulo  $p^n$ .*

**PROOF.** We will proceed by induction on  $n$ . In the case of  $n = 1$ , the congruence  $f(x) \equiv 0 \pmod{p^1}$  is only another formulation of the assumption that the integer  $a$  satisfies  $p \mid f(a)$ . Further, let  $n > 1$  and suppose the proposition is true for  $n - 1$ . If  $x$  satisfies the system for  $n$ , then it does so for  $n - 1$  as well. Denoting one of the solutions of the system for  $n - 1$  as  $c_{n-1}$ , we can look for the solution of the system for  $n$  in the form

$$x = c_{n-1} + k \cdot p^{n-1}, \quad \text{where } k \in \mathbb{Z}.$$

We need to find out for which  $k$  we have  $f(c_{n-1} + k \cdot p^{n-1}) \equiv 0 \pmod{p^n}$ . We know that  $p^{n-1} \mid f(c_{n-1} + k \cdot p^{n-1})$ . Now, we use the binomial theorem for  $f(x) = a_m x^m + \dots + a_1 x + a_0$ , where  $a_0, \dots, a_m \in \mathbb{Z}$ . We have

$$(c_{n-1} + k \cdot p^{n-1})^i \equiv c_{n-1}^i + i \cdot c_{n-1}^{i-1} \cdot k p^{n-1} \pmod{p^n},$$

hence

$$f(c_{n-1} + k \cdot p^{n-1}) \equiv f(c_{n-1}) + k \cdot p^{n-1} f'(c_{n-1}).$$

Therefore,

$$f(c_{n-1} + k \cdot p^{n-1}) \equiv 0 \pmod{p^n} \iff$$

$$\iff 0 \equiv \frac{f(c_{n-1})}{p^{n-1}} + k \cdot f'(c_{n-1}) \pmod{p}.$$

Since  $c_{n-1} \equiv a \pmod{p}$ , we get  $f'(c_{n-1}) \equiv f'(a) \not\equiv 0 \pmod{p}$ , so  $(f'(c_{n-1}), p) = 1$ . By the theorem about the solutions of linear congruences, we can hence see that there is (modulo  $p$ ) a unique solution  $k$  of this congruence, and since  $c_{n-1}$  was, by the induction hypothesis, the only solution modulo  $p^{n-1}$ , the integer  $c_{n-1} + k \cdot p^{n-1}$  is the only solution of the given system modulo  $p^n$ .  $\square$

**Example.** Consider the congruence

$$3x^2 + 4 \equiv 0 \pmod{49}.$$

The congruence can be equivalently transformed (by solving the linear congruence  $3y \equiv 1 \pmod{49}$  and multiplying both sides of the congruence by the integer  $y \equiv 33$ ) to the form  $x^2 \equiv 15 \pmod{7^2}$ . Then, we proceed as in the constructive proof of Hensel’s lemma.

and if not, it must be a quadratic residue modulo 43. The quadratic residues can be most simply enumerated by calculating the squares of all elements of a reduced residue system modulo 43.

The quadratic residues are thus the integers congruent to  $(\pm 1)^2, (\pm 2)^2, (\pm 3)^2, \dots, (\pm 21)^2$  modulo 43, so the problem is satisfied by exactly those integers  $a$  which are congruent to any one of 1, 4, 6, 9, 10, 11, 13, 14, 15, 16, 17, 21, 23, 24, 25, 31, 35, 36, 38, 40, 41.  $\square$

Using law of quadratic reciprocity 11.4.13 we can calculate the value  $(a/p)$  for any integer  $a$  and an odd prime  $p$ . Moreover, evaluation of the Legendre symbol is fast enough even for high arguments, therefore using it is favourable to verifying criteria of the theorem 11.4.10.

**11.D.27.** Here, we recall the statement of the Law in the slightly modified way which is more suitable for direct calculations.

- i)  $-1$  is a quadratic residue for primes  $p$  which satisfy  $p \equiv 1 \pmod{4}$  and it is a quadratic nonresidue for primes  $p$  satisfying  $p \equiv 3 \pmod{4}$ .
- ii)  $2$  is a quadratic residue for primes  $p$  which satisfy  $p \equiv \pm 1 \pmod{8}$  and it is a quadratic nonresidue for primes  $p$  satisfying  $p \equiv \pm 3 \pmod{8}$ .
- iii) If  $p \equiv 1 \pmod{4}$  or  $q \equiv 1 \pmod{4}$ , then  $(p/q) = (q/p)$ ; for other odd  $p, q$ , we have  $(p/q) = -(q/p)$ .

**Solution.** We simply apply law of quadratic reciprocity in the appropriate cases.

- i) The integer  $\frac{p-1}{2}$  is even iff  $4 \mid p - 1$ .
- ii) We need to know for which odd primes  $p$  the exponent is  $\frac{p^2-1}{8}$  is even. Odd primes are congruent to  $\pm 1$  or  $\pm 3$  modulo 8, so we have (by 11.C.7) that either  $p^2 \equiv 1 \pmod{16}$  or  $p^2 \equiv 9 \pmod{16}$ .  $\square$
- iii) This is clear from the law of quadratic reciprocity.

**11.D.28.** Derive by straight calculation from Gauss's lemma 11.4.14 once again the so-called supplementary laws of quadratic reciprocity:

$$\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}} \quad \text{and} \quad \left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}.$$

**Solution.** To evaluate  $(-1/p)$  in the former case, we should realize that  $\mu$  tells the number of least (in absolute value) negative remainders of integers in the set

$$\{-1, -2, \dots, -\frac{p-1}{2}\}.$$

However, those are exactly the desired remainders and they are all negative; hence we have  $\mu = \frac{p-1}{2}$  and  $(-1/p) = (-1)^{\frac{p-1}{2}}$ .

In the latter case, we need to express the number of least (in absolute value) negative remainders of integers in the set

$$\{1 \cdot 2, 2 \cdot 2, 3 \cdot 2, \dots, \frac{p-1}{2} \cdot 2\}.$$

First, we solve the congruence  $x^2 \equiv 15 \equiv 1 \pmod{7}$ , which has at most 2 solutions, and those are  $x \equiv \pm 1 \pmod{7}$ . These solutions can be expressed in the form  $x = \pm 1 + 7t$ , where  $t \in \mathbb{Z}$ , and substituted into the congruence modulo 49, whence we get the solution  $x \equiv \pm 8 \pmod{49}$  (if we were interested solely in the number of solutions, we would not even have to finish the calculation as it follows straight from Hensel's lemma that every solution modulo 7 gives a unique solution modulo 49 because for  $f(x) = x^2 - 15$ , we have  $7 \nmid f'(\pm 1)$ ).

**11.4.7. Congruences modulo a prime.** The solution of general higher-order congruences has thus been reduced to the solution of congruences modulo a prime. As we will see, this is where the stumbling block is since no (much) more efficient universal procedure than trying out all possibilities is known. We can at least mention several statements describing the solvability and number of solutions of such congruences. We will then prove some detailed results for some special cases in further paragraphs.



**Theorem.** Let  $p$  be a prime,  $f(x) \in \mathbb{Z}[x]$ . Every congruence  $f(x) \equiv 0 \pmod{p}$  is equivalent to a congruence of degree at most  $p - 1$ .

**PROOF.** Since it holds for any  $a \in \mathbb{Z}$  that  $p \mid a^p - a$  (simple consequence of Fermat's little theorem), the congruence  $x^p - x \equiv 0 \pmod{p}$  is satisfied by all integers. Dividing the polynomial  $f(x)$  by  $x^p - x$  with remainder, we get

$$f(x) = q(x) \cdot (x^p - x) + r(x)$$

for suitable  $f(x), r(x) \in \mathbb{Z}$ , where the degree of  $r(x)$  is less than that of the divisor, i.e.  $p$ . We thus get that the congruence  $r(x) \equiv 0 \pmod{p}$  is equivalent to the congruence  $f(x) \equiv 0 \pmod{p}$ , yet it is of degree at most  $p - 1$ .  $\square$

**11.4.8. Theorem.** Let  $p$  be a prime,  $f(x) \in \mathbb{Z}[x]$ . If the congruence  $f(x) \equiv 0 \pmod{p}$  has more than  $\deg(f)$  solutions, then each of the coefficients of the polynomial  $f$  is a multiple of  $p$ .

**PROOF.** In algebraic words, we are actually interested in the number of roots of a non-zero polynomial over a finite field  $\mathbb{Z}_p$ , and by 12.3.4, there are at most  $\deg(f)$  of them.  $\square$

**Corollary** (Another proof of Wilson's theorem). If  $p$  is a prime, then

$$(p - 1)! \equiv -1 \pmod{p}.$$

**PROOF.** The statement is clearly true for  $p = 2$ , so we can consider only odd primes  $p$  from now on. By Fermat's little theorem, the congruence

$$(x - 1)(x - 2) \cdots (x - (p - 1)) - (x^{p-1} - 1) \equiv 0 \pmod{p}$$

is satisfied by any integer  $a$  which is not divisible by  $p$ ; i.e., there are  $p - 1$  solutions. However, its degree is equal to  $p - 2$  (which is less than the number of solutions). It follows from 11.4.7 that all of the coefficients of the left-hand polynomial

For any  $k \in \{1, 2, \dots, \frac{p-1}{2}\}$ , the integer  $2k$  leaves a negative remainder if and only if  $2k > \frac{p-1}{2}$ , i.e., iff  $k > \frac{p-1}{4}$ . Now, it remains to determine the number of such integers  $k$ .

If  $p \equiv 1 \pmod{4}$ , then this number is equal to  $\frac{p-1}{2} - \frac{p-1}{4} = \frac{p-1}{4}$ , so

$$\left(\frac{-1}{p}\right) = (-1)^\mu = (-1)^{\frac{p-1}{4}} = (-1)^{\frac{p-1}{4} \cdot \frac{p+1}{2}} = (-1)^{\frac{p^2-1}{8}},$$

since  $\frac{p+1}{2}$  is odd in this case.

Similarly, for  $p \equiv 3 \pmod{4}$ , the number of such integers  $k$  equals  $\frac{p-1}{2} - \frac{p-3}{4} = \frac{p+1}{4}$ , so

$$\left(\frac{-1}{p}\right) = (-1)^{\frac{p+1}{4}} = (-1)^{\frac{p+1}{4} \cdot \frac{p-1}{2}} = (-1)^{\frac{p^2-1}{8}},$$

since  $\frac{p-1}{2}$  is odd in this case as well. □

**11.D.29.** Solve the congruence  $x^2 - 23 \equiv 0 \pmod{77}$ .

**Solution.** Factoring the modulus, we get the system

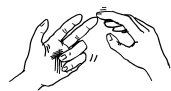
$$\begin{aligned} x^2 - 1 &\equiv 0 \pmod{11}, \\ x^2 - 2 &\equiv 0 \pmod{7}. \end{aligned}$$

Clearly, 1 is a quadratic residue modulo 11, so the first congruence of the system has (exactly) two solutions:  $x \equiv \pm 1 \pmod{11}$ . Further,  $(2/7) = (9/7) = 1$ , and it should not take much effort to notice the solution:  $x \equiv \pm 3 \pmod{7}$ .

We have thus obtained four simple systems of two linear congruences each. Solving them, we will get that the original congruence has the following four solutions:  $x \equiv 10, 32, 45$  or  $67 \pmod{77}$ . □

**11.D.30.** Solve the congruence  $7x^2 + 112x + 42 \equiv 0 \pmod{473}$ . ○

**Jacobi symbol.** Jacobi symbol  $(a/b)$  is a generalization of the Legendre symbol to the case where the “lower” argument  $b$  need not be a prime, but any odd positive integer. It is defined as the product of the Legendre symbols corresponding to the prime factors of  $b$ : if  $b = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ , then



$$\left(\frac{a}{b}\right) = \left(\frac{a}{p_1}\right)^{\alpha_1} \cdots \left(\frac{a}{p_k}\right)^{\alpha_k}.$$

The primary motivation for introducing the Jacobi symbol is the necessity to evaluate the Legendre symbol (and thus to decide the solvability of quadratic congruences) without having to factor integers to primes. We will illustrate such calculation on an example having in mind that Jacobi symbol shares with the Legendre one not only the notation but also almost all of the (computational) properties.

**11.D.31.** Decide whether the congruence  $x^2 \equiv 219 \pmod{383}$  is solvable.

are multiples of  $p$ . In particular, this applies to the absolute term, which equals  $(p-1)!+1$ . This proves Wilson’s theorem. □

**11.4.9. Binomial congruences.** This part will be devoted to solving special types of higher-order polynomial congruences, the so-called *binomial congruences*. It is an analog to the binomial equations, where the polynomial  $f(x)$  is  $x^n - a$ . It can easily be shown that we can restrict ourselves to the condition that  $a$  be coprime with the modulus of the congruence – otherwise, we can always equivalently transform the congruence into this form or decide that it has no solution.



QUADRATIC AND POWER RESIDUES

Let  $m \in \mathbb{N}$ ,  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ . The integer  $a$  is said to be a *n-th power residue modulo m*, or *residue of degree n modulo m* if and only if the congruence

$$x^n \equiv a \pmod{m}$$

is solvable. Otherwise, we call  $a$  a *n-th power nonresidue modulo m*, or *nonresidue of degree n modulo m*.

For  $n = 2, 3, 4$ , we use the adjectives quadratic, cubic, and quartic residue (or nonresidue) modulo  $m$ .

Now, we will show how to solve binomial congruences modulo  $m$ , if there are primitive roots modulo  $m$  (in particular, when the modulus is an odd prime or its power).

**11.4.10. Theorem.** Let  $m \in \mathbb{N}$  be such that there are primitive roots modulo  $m$ . Further, let  $a \in \mathbb{Z}$ ,  $(a, m) = 1$ . Then, the congruence  $x^n \equiv a \pmod{m}$  is solvable (i.e.,  $a$  is an *n-th power residue modulo m*) if and only if  $a^{\varphi(m)/d} \equiv 1 \pmod{m}$ , where  $d = (n, \varphi(m))$ . And if so, it has exactly  $d$  solutions.

**PROOF.** Let  $g$  be a primitive root modulo  $m$ . Then, for any  $x$  coprime to  $m$ , there is a unique integer  $y$  (its discrete logarithm) with the property  $0 \leq y < \varphi(m)$  such that  $x \equiv g^y \pmod{m}$ . Similarly, for a given  $a$ , there is a unique  $b \in \mathbb{Z}$ ;  $0 \leq b < \varphi(m)$  such that  $a \equiv g^b \pmod{m}$ . After this substitution, the binomial congruence in question is thus equivalent to the congruence  $(g^y)^n \equiv g^b \pmod{m}$  and, invoking theorem 11.3.13, to the linear congruence  $n \cdot y \equiv b \pmod{\varphi(m)}$  as well.

However, this congruence

$$n \cdot y \equiv b \pmod{\varphi(m)}$$

is solvable if and only if  $d = (n, \varphi(m)) \mid b$  (and if so, it has  $d$  solutions).

It remains to prove that  $d \mid b$  if and only if  $a^{\varphi(m)/d} \equiv 1 \pmod{m}$ . However, the congruence

$$1 \equiv a^{\varphi(m)/d} \equiv g^{b\varphi(m)/d} \pmod{m}$$

is true if and only if  $\varphi(m) \mid \frac{b\varphi(m)}{d}$ , which happens if and only if  $d \mid b$ . □

**Solution.** Since 383 is a prime, the congruence will be solvable if the Legendre symbol will satisfy  $(219/383) = 1$ .

$$\begin{aligned} \left(\frac{219}{383}\right) &= -\left(\frac{383}{219}\right) \text{ (Jacobi) as } 383 \equiv 219 \equiv 3 \pmod{4} \\ &= -\left(\frac{164}{219}\right) = -\left(\frac{41}{219}\right) \quad 164 = 2^2 \cdot 41 \\ &= -\left(\frac{219}{41}\right) \text{ (Jacobi) as } 41 \equiv 1 \pmod{4} \\ &= -\left(\frac{14}{41}\right) = -\left(\frac{2}{41}\right)\left(\frac{7}{41}\right) \\ &= -\left(\frac{7}{41}\right) \quad \text{as } 41 \equiv 1 \pmod{8} \\ &= -\left(\frac{41}{7}\right) \quad \text{as } 41 \equiv 1 \pmod{4} \\ &= -\left(\frac{-1}{7}\right) = 1 \quad \text{as } 7 \equiv 3 \pmod{4}. \end{aligned}$$

□

Now, we introduce several exercises proving that the Jacobi symbol has properties similar to the Legendre one, which relieves us of the necessity to factor the integers that appear when working purely with the Legendre symbol.

**11.D.32.** Prove that all odd positive numbers  $b, b'$  and all integers  $a, a_1, a_2$  satisfy (the symbols used here are always the Jacobi ones):

- i) if  $a_1 \equiv a_2 \pmod{b}$ , then  $\left(\frac{a_1}{b}\right) = \left(\frac{a_2}{b}\right)$ ,
- ii)  $\left(\frac{a_1 a_2}{b}\right) = \left(\frac{a_1}{b}\right)\left(\frac{a_2}{b}\right)$ ,
- iii)  $\left(\frac{a}{bb'}\right) = \left(\frac{a}{b}\right)\left(\frac{a}{b'}\right)$ .

○

**11.D.33.** Prove that if  $a, b$  are odd natural numbers, then

- i)  $\frac{ab-1}{2} \equiv \frac{a-1}{2} + \frac{b-1}{2} \pmod{2}$ ,
- ii)  $\frac{a^2 b^2 - 1}{8} \equiv \frac{a^2 - 1}{8} + \frac{b^2 - 1}{8} \pmod{2}$ .

**Solution.**

- i) Since the integer  $(a-1)(b-1) = (ab-1) - (a-1) - (b-1)$  is a multiple of 4, we get  $(ab-1) \equiv (a-1) + (b-1) \pmod{4}$ , which gives what we want when divided by two.
- ii) Similarly to above,  $(a^2-1)(b^2-1) = (a^2 b^2 - 1) - (a^2 - 1) - (b^2 - 1)$  is a multiple of 16. Therefore,  $(a^2 b^2 - 1) \equiv (a^2 - 1) + (b^2 - 1) \pmod{16}$ , which gives the wanted statement when divided by eight (see also exercise 11.A.2). □

**11.D.34.** Prove that if  $a_1, \dots, a_k$  are odd natural numbers, then

- i)  $\prod_{\ell=1}^k \frac{a_\ell - 1}{2} \equiv \sum_{\ell=1}^k \frac{a_\ell - 1}{2} \pmod{2}$ ,
- ii)  $\prod_{\ell=1}^k \frac{a_\ell^2 - 1}{8} \equiv \sum_{\ell=1}^k \frac{a_\ell^2 - 1}{8} \pmod{2}$ .

**Corollary.** If the assumptions of the above theorem hold and, moreover,  $(n, \varphi(m)) = 1$ , the congruence  $x^n \equiv a \pmod{m}$  always has a unique solution. In other words, exponentiation to the  $n$ -th power (where  $n$  is coprime to  $\varphi(m)$ ) is a bijection on the set  $\mathbb{Z}_m^\times$  of invertible residue classes modulo  $m$  (it is even an automorphism of the group  $(\mathbb{Z}_m^\times, \cdot)$ ).

**11.4.11. Quadratic congruences and the Legendre symbol.** Now, our task is to find an efficient condition determining whether a quadratic congruence



$$ax^2 + bx + c \equiv 0 \pmod{m}$$

is solvable (and if so, how many solutions it has). It can easily be seen from the presented theory that if we want to decide whether this congruence is solvable, it suffices to decide this for the (binomial) congruence

$$x^2 \equiv a \pmod{p},$$

where  $p$  is an odd prime and  $a$  is an integer coprime to it. A congruence modulo a composite  $m$  can be decomposed to an equivalent system of congruences modulo the particular factors of the integer  $m$ , which are powers of primes. Such congruences can be transformed to quadratic congruences with prime modulus using the procedure described in Hensel's lemma 11.4.6. Norming this congruence and completing the square then results in the aforementioned form.

To decide the solvability of a congruence, we can, of course, use the theorem 11.4.10 about the solvability of binomial congruences. Its application is, however, often limited by time resources; we will thus try to find a criterion which will be computationally easier in (not only) the quadratic case.

**Example.** Let us determine the number of solutions of the congruence  $x^2 \equiv 219 \pmod{383}$ .

Since 383 is a prime and  $(2, \varphi(383)) = 2$ , it follows from theorem 11.4.10 that the given congruence is solvable (and it has 2 solutions) if and only if  $219^{\frac{\varphi(383)}{2}} = 219^{191} \equiv 1 \pmod{383}$ . It is not easy to verify this proposition without some computational power (though, this can still be calculated on a "piece of paper"). However, we will show that this condition can be verified much more easily using the properties of the so-called Legendre symbol.

#### LEGENDRE SYMBOL

Let  $p$  be an odd prime and  $a$  an integer. The Legendre symbol is defined by

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & \text{for } p \nmid a, a \text{ is a quadratic residue modulo } p, \\ 0 & \text{for } p \mid a, \\ -1 & \text{if } a \text{ is a quadratic nonresidue modulo } p. \end{cases}$$

The Legendre symbol is also often written as  $(a/p)$  and usually read as "a on p".

**Example.** Since the congruence  $x^2 \equiv 1 \pmod{p}$  is solvable for an arbitrary odd prime  $p$ , we have  $(1/p) = 1$ . Further,

**11.D.35.** Prove the law of quadratic reciprocity for the Jacobi symbol, i.e., prove that if  $a, b$  are odd natural numbers, then

- i)  $\left(\frac{-1}{a}\right) = (-1)^{\frac{a-1}{2}}$ ,
- ii)  $\left(\frac{2}{a}\right) = (-1)^{\frac{a^2-1}{8}}$ ,
- iii)  $\left(\frac{a}{b}\right) = \left(\frac{b}{a}\right) \cdot (-1)^{\frac{a-1}{2} \frac{b-1}{2}}$ .

**Solution.** Let (just like in the definition of the Jacobi symbol)  $a$  factor to (odd) primes as  $p_1 p_2 \cdots p_k$ .

- i) The properties of the Legendre symbol and the aforementioned statement imply that

$$\begin{aligned} \left(\frac{-1}{a}\right) &= \left(\frac{-1}{p_1}\right) \cdot \left(\frac{-1}{p_2}\right) \cdots \left(\frac{-1}{p_k}\right) = \\ &= (-1)^{\frac{p_1-1}{2}} \cdots (-1)^{\frac{p_k-1}{2}} = \\ &= (-1)^{\sum_{i=1}^k \frac{p_i-1}{2}} = \\ &= (-1)^{\prod_{i=1}^k \frac{p_i-1}{2}} = (-1)^{\frac{a-1}{2}}. \end{aligned}$$

- ii) Analogously to above.
- iii) Further, let  $b$  factor to (odd) primes as  $q_1 q_2 \cdots q_\ell$ . If we have  $p_i = q_j$  for some  $i$  and  $j$ , then the symbols on both sides of the equality are equal to zero. Otherwise, the law of quadratic reciprocity for the Legendre symbol implies that for all pairs  $(p_i, q_j)$ , we have

$$\left(\frac{p_i}{q_j}\right) = \left(\frac{q_j}{p_i}\right) \cdot (-1)^{\frac{p_i-1}{2} \frac{q_j-1}{2}}.$$

Therefore,

$$\begin{aligned} \left(\frac{a}{b}\right) &= \prod_{i=1}^k \prod_{j=1}^{\ell} \left(\frac{p_i}{q_j}\right) = \\ &= \prod_{i=1}^k \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) \cdot (-1)^{\frac{p_i-1}{2} \frac{q_j-1}{2}} = \\ &= \prod_{i=1}^k (-1)^{\frac{p_i-1}{2} \sum_{j=1}^{\ell} \frac{q_j-1}{2}} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) = \\ &= \prod_{i=1}^k (-1)^{\frac{p_i-1}{2} \frac{\prod_{j=1}^{\ell} q_j-1}{2}} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) = \\ &= \prod_{i=1}^k (-1)^{\frac{p_i-1}{2} \frac{b-1}{2}} \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) = \\ &= (-1)^{\frac{b-1}{2} \sum_{i=1}^k \frac{p_i-1}{2}} \prod_{i=1}^k \prod_{j=1}^{\ell} \left(\frac{q_j}{p_i}\right) = \\ &= (-1)^{\frac{a-1}{2} \frac{b-1}{2}} \left(\frac{b}{a}\right). \end{aligned}$$

We utilized the result of part (i) of the previous exercise in the calculations.  $\square$

**11.D.36.** Determine whether the congruence  $x^2 \equiv 38 \pmod{165}$  is solvable.

$(-1/5) = (4/5) = 1$ , because the congruence  $x^2 \equiv -1 \pmod{5}$  is equivalent to the congruence  $x^2 \equiv 4 \pmod{5}$ , whose solutions are  $x \equiv \pm 2 \pmod{5}$ .

The statement of the following lemma will be very often used when evaluating the Legendre symbol in practice.

**11.4.12. Lemma.** Let  $p$  be an odd prime,  $a, b \in \mathbb{Z}$  arbitrary. Then:

- (1)  $\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}$ .
- (2)  $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right) \left(\frac{b}{p}\right)$ .
- (3) If  $a \equiv b \pmod{p}$ , then  $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$ .

**PROOF.** (1) The statement is clear for  $p \mid a$ ; if  $a$  is a quadratic residue modulo  $p$ , then the statement follows from the theorem about the solvability of quadratic congruences, which claims (in this case, we have  $(\varphi(p), 2) = 2$ ) that the necessary and sufficient condition for the congruence  $x^2 \equiv a \pmod{p}$  to be solvable that

$$a^{\frac{p-1}{2}} \equiv 1 \pmod{p}.$$

The same theorem implies for the case of a quadratic non-residue as well that we have  $a^{\frac{p-1}{2}} \not\equiv 1 \pmod{p}$ . However, then (since we have  $p \mid a^{p-1} - 1 = (a^{\frac{p-1}{2}} - 1)(a^{\frac{p-1}{2}} + 1)$  by Fermat's theorem), necessarily  $p \mid a^{\frac{p-1}{2}} + 1$ , i.e.,  $a^{\frac{p-1}{2}} \equiv -1 \pmod{p}$ .

- (2) From (1), we have

$$\left(\frac{ab}{p}\right) \equiv (ab)^{\frac{p-1}{2}} = a^{\frac{p-1}{2}} b^{\frac{p-1}{2}} \equiv \left(\frac{a}{p}\right) \left(\frac{b}{p}\right) \pmod{p}.$$

However, since the values of the Legendre symbol belong to the set  $\{-1, 0, 1\}$ , this congruence immediately implies that the left and right sides are equal.

- (3) Apparent from the definition.  $\square$

**Corollary.** (1) Any reduced residue system modulo  $p$  contains the same number of quadratic residues as non-residues.

- (2) The product of two quadratic residues as well as the product of two quadratic nonresidues is a residue; the product of a residue and a nonresidue is a nonresidue.
- (3)  $(-1/p) = (-1)^{\frac{p-1}{2}}$ , i.e., the congruence  $x^2 \equiv -1 \pmod{p}$  is solvable if and only if  $p \equiv 1 \pmod{4}$ .

**PROOF.** (1) Considering the elements of a reduced residue system modulo  $p$  (we can take, for instance, the set  $\{-\frac{p-1}{2}, \dots, -1, 1, \dots, \frac{p-1}{2}\}$ ), the quadratic residues are exactly those integers which are congruent to one of  $(\pm 1)^2, \dots, (\pm \frac{p-1}{2})^2$ . Thus there are exactly  $\frac{p-1}{2}$  of quadratic residues, so there are  $p-1 - \frac{p-1}{2} = \frac{p-1}{2}$  of the other ones (the quadratic nonresidues).

- (2) This follows immediately from part (2) and the previous lemma.

**Solution.** The Jacobi symbol is equal to

$$\left(\frac{38}{165}\right) = \left(\frac{2}{165}\right) \cdot \left(\frac{19}{165}\right) = \left(\frac{2}{3}\right) \cdot \left(\frac{2}{5}\right) \cdot \left(\frac{2}{11}\right) \cdot \left(\frac{19}{3}\right) \cdot \left(\frac{19}{5}\right) \cdot \left(\frac{19}{11}\right) = (-1)^3 \left(\frac{1}{3}\right) \cdot \left(\frac{-1}{5}\right) \cdot \left(\frac{2}{11}\right)^3 = 1.$$

This result does not answer the question of the existence of a solution. However, if we split the congruence to a system of congruences according to the factors of the modulus, we obtain

$$\begin{aligned} x^2 &\equiv -1 \pmod{3}, \\ x^2 &\equiv 3 \pmod{5}, \\ x^2 &\equiv 5 \pmod{11}, \end{aligned}$$

whence we can easily see that the first and second congruences have no solution. In particular,

$$\left(\frac{-1}{3}\right) = -1 \quad \text{and} \quad \left(\frac{3}{5}\right) = \left(\frac{5}{3}\right) = \left(\frac{2}{3}\right) = -1.$$

Therefore, neither the original congruence has a solution.  $\square$

**11.D.37.** Find all primes  $p$  such that the integer below is a quadratic residue modulo  $p$ :



- i) 3,    ii)  $-3$ ,    iii) 6.

**Solution.**

- i) We are looking for primes  $p \neq 3$  such that  $x^2 \equiv 3 \pmod{p}$  is solvable. Since  $p = 2$  satisfies the above, we will consider only odd primes  $p \neq 3$  from now on. For  $p \equiv 1 \pmod{4}$ , it follows from the law of quadratic reciprocity that  $1 = (3/p) = (p/3)$ , which occurs if and only if  $p \equiv 1 \pmod{3}$ . On the other hand, if  $p \equiv -1 \pmod{4}$ , then  $1 = (3/p) = -(p/3)$ , which holds for  $p \equiv -1 \pmod{3}$ . Putting the conditions of both cases together, we arrive at  $p \equiv \pm 1 \pmod{12}$ , which, together with  $p = 2$ , completes the set of all primes satisfying the given condition.
- ii) The condition  $1 = (-3/p) = (-1/p)(3/p)$  is satisfied if either  $(-1/p) = (3/p) = 1$  or  $(-1/p) = (3/p) = -1$ . In the former case (using the result of the previous item), this means that  $p \equiv 1 \pmod{4}$  and  $p \equiv \pm 1 \pmod{12}$ . In the latter case, we must have  $p \equiv -1 \pmod{4}$  and  $p \equiv \pm 5 \pmod{12}$ , at the same time – we can take, for instance, the set  $\{-5, -1, 1, 5\}$  for a reduced residue system modulo 12, and since  $(3/p) = 1$  for  $p \equiv \pm 1 \pmod{12}$ , we surely have  $(3/p) = -1$  whenever  $p \equiv \pm 5 \pmod{12}$ . We have thus obtained four systems of two congruences each. Two of them have no solution, and the remaining two are satisfied by  $p \equiv 1 \pmod{12}$  and  $p \equiv -5 \pmod{12}$ , respectively.
- iii) In this case,  $(6/p) = (2/p)(3/p)$  and once again, there are two possibilities: either  $(2/p) = (3/p) = 1$  or  $(2/p) = (3/p) = -1$ . The former case occurs if  $p$  satisfies  $p \equiv \pm 1 \pmod{8}$  as well as  $p \equiv \pm 1 \pmod{12}$ . Solving the corresponding systems of linear congruences leads to the condition  $p \equiv \pm 1 \pmod{24}$ . In the latter case, we get  $p \equiv \pm 3 \pmod{8}$  as well as  $p \equiv \pm 5 \pmod{12}$ , which together gives  $p \equiv \pm 5 \pmod{24}$ .

Let us remark that thanks to Dirichlet's theorem 11.2.5, the number of primes we were interested in is infinite in each of the three problems.  $\square$

- (3) It follows from part (1) of the lemma that  $(-1/p) \equiv (-1)^{\frac{p-1}{2}} \pmod{p}$ ; both sides, however, take on the values  $\pm 1$ , so they must be equal.  $\square$

These basic statements about the values of the Legendre symbol are already sufficient for proving the theorem on the infinitude of primes of the form  $4k + 1$  (see paragraph 11.2.3).



**Proposition.** *There are infinitely many primes of the form  $4k + 1$ .*

**PROOF.** We will proceed by contradiction. Suppose that  $p_1, p_2, \dots, p_\ell$  is the enumeration of all primes of the form  $4k + 1$ , and consider the integer  $N = (2p_1 \cdots p_\ell)^2 + 1$ . This integer is of the form  $4k + 1$  as well. The assumption that  $N$  is a prime would lead to an immediate contradiction, since  $N$  is surely greater than any of the integers  $p_1, p_2, \dots, p_\ell$ . Therefore, from now on, let us suppose that it is thus composite. Then, there must exist a prime  $p$  which divides  $N$ . Apparently, none of the primes  $2, p_1, p_2, \dots, p_\ell$  divides  $N$ , so we will be finished if we prove that  $p$  is also of the form  $4k + 1$ . It follows from the congruence  $(2p_1 \cdots p_\ell)^2 \equiv -1 \pmod{p}$ , that  $(-1/p) = 1$ , and this is true (by the previous corollary) if and only if  $p \equiv 1 \pmod{4}$ . Altogether, we have reached a contradiction (a prime  $p$  not belonging to the original list of all primes of the form  $4k + 1$ ) in the case of composite  $N$  as well, which proves that there are infinitely many such primes.  $\square$

The most important theorem which allows us to efficiently compute the value of the Legendre symbol (and thus determine the solvability of a quadratic congruence), is the so-called *law of quadratic reciprocity*.



#### LAW OF QUADRATIC RECIPROCITY

**11.4.13. Theorem.** *Let  $p, q$  be odd primes. Then,*

- (1)  $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}$ ,
- (2)  $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$ ,
- (3)  $\left(\frac{q}{p}\right) = \left(\frac{p}{q}\right) \cdot (-1)^{\frac{p-1}{2} \cdot \frac{q-1}{2}}$ .

The theorem is put this way mainly because we can calculate the value  $(a/p)$  for any integer  $a$  using these three formulae and the basic rules for the Legendre symbol.

**Example.** Let us calculate the value  $(79/101)$  using the properties of the Legendre symbol.



**11.D.38.** The following exercise illustrates that if the modulus of a quadratic congruence is a prime  $p$  satisfying  $p \equiv 3 \pmod{4}$ , then we are able not only to decide the solvability of the congruence, but also to describe all of its solutions in a simple way.

Consider a prime  $p \equiv 3 \pmod{4}$  and an integer  $a$  such that  $(a/p) = 1$ . Prove that the solution of the congruence  $x^2 \equiv a \pmod{p}$  is

$$x \equiv \pm a^{\frac{p+1}{4}} \pmod{p}.$$

**Solution.** It can be easily verified (using lemma 11.4.12) that  $(a^{\frac{p+1}{4}})^2 \equiv a^{\frac{p+1}{2}} \equiv a \cdot \left(\frac{a}{p}\right) \equiv a \pmod{p}$ .  $\square$

**11.D.39.** Determine whether the congruence

$$x^2 \equiv 3 \pmod{59}$$

is solvable. If so, find all of its solutions.

**Solution.** Calculating the Legendre symbol

$$\left(\frac{3}{59}\right) = -\left(\frac{59}{3}\right) = -\left(\frac{2}{3}\right) = -(-1) = 1,$$

we find out that the congruence has two solutions. Thanks to the statement above, we can immediately see  $(59 \equiv 3 \pmod{4})$  that the congruence is satisfied by

$$\begin{aligned} x &\equiv \pm 3^{\frac{59+1}{4}} = \pm 3^{15} \equiv (3^5)^3 \equiv \\ &\equiv \pm 7^3 = \pm 343 \equiv \mp 11 \pmod{59}, \end{aligned}$$

since  $3^5 = 243 \equiv 7 \pmod{59}$ .  $\square$

### E. Diophantine equations

Here, we limit ourselves only to the small class of equations which can be solved using divisibility or can be reduced to solving congruences.

**11.E.1. Linear Diophantine equations.** Decide whether it is possible to use a balance scale to weigh 50 grams of given goods provided we have only (an arbitrary number of) three kinds of masses; their weights are 770, 630, and 330 grams, respectively.

If so, how to do that?

**Solution.** Our task is to solve the equation

$$770x + 630y + 330z = 50,$$

where  $x, y, z \in \mathbb{Z}$  (a negative value in the solution would mean that we put the corresponding masses on the other scale). Dividing both sides of the equation by  $(770, 630, 330) = 10$ , we get an equivalent equation

$$77x + 63y + 33z = 5.$$

Considering this equation modulo  $(77, 63) = 7$ , we get the following linear congruence:

$$\begin{aligned} 33z &\equiv 5 \pmod{7}, \\ 5z &\equiv 5 \pmod{7}, \\ z &\equiv 1 \pmod{7}. \end{aligned}$$



$$\begin{aligned} \left(\frac{79}{101}\right) &= \left(\frac{101}{79}\right) \quad \text{since 101 is congruent to 1 modulo 4} \\ &= \left(\frac{22}{79}\right) \\ &= \left(\frac{2}{79}\right) \cdot \left(\frac{11}{79}\right) \\ &= \left(\frac{11}{79}\right) \quad \text{since 79 is congruent to } -1 \text{ modulo 8} \\ &= (-1) \left(\frac{79}{11}\right) \quad \text{since } 11 \equiv 79 \equiv 3 \pmod{4} \\ &= (-1) \left(\frac{2}{11}\right) = 1 \quad \text{since } 11 \equiv 3 \pmod{8}. \end{aligned}$$

Many proofs of the the quadratic reciprocity law can be found in literature<sup>6</sup>. However, many of them (especially the shorter ones) usually make use of deeper knowledge from algebraic number theory. We will present an elementary proof of this theorem here.

Let  $S$  denote the reduced residue system of the least residues (in absolute value) modulo  $p$ , i.e.,

$$S = \left\{ -\frac{p-1}{2}, -\frac{p-3}{2}, \dots, -1, 1, \dots, \frac{p-3}{2}, \frac{p-1}{2} \right\}.$$

Further, for  $a \in \mathbb{Z}$ ,  $p \nmid a$ , let  $\mu_p(a)$  denote the number of negative least residues (in absolute value) of the integers

$$1 \cdot a, 2 \cdot a, \dots, \frac{p-1}{2} \cdot a,$$

i.e., we decide for each of these integers to which integer from the set  $S$  it is congruent and count the number of the negative ones. If it is clear from context which values  $a, p$  we mean, we will usually omit the parameters and write only  $\mu$  instead of  $\mu_p(a)$ .

**Example.** We determine  $\mu_p(a)$  for the prime  $p = 11$  and the integer  $a = 3$ .

Now, the reduced residue system we are interested in is  $S = \{-5, \dots, -1, 1, \dots, 5\}$ , and for  $a = 3$ , we calculate

$$\begin{aligned} 1 \cdot 3 &\equiv 3 \pmod{11} \\ 2 \cdot 3 &\equiv -5 \pmod{11} \\ 3 \cdot 3 &\equiv -2 \pmod{11} \\ 4 \cdot 3 &\equiv 1 \pmod{11} \\ 5 \cdot 3 &\equiv 4 \pmod{11}, \end{aligned}$$

whence  $\mu_{11}(3) = 2$ .

We will show in the following statement that this integer is tightly connected to the Legendre symbol – the value of the symbol  $(3/11)$  can be determined in terms of the  $\mu$  function as  $(-1)^{\mu_{11}(3)} = (-1)^2 = 1$ .

<sup>6</sup>In 2000, F. Lemmermeyer stated 233 proofs – see F. Lemmermeyer, *Reciprocity laws. From Euler to Eisenstein*, Springer, 2000

This congruence is thus satisfied by those integers  $z$  of the form  $z = 1 + 7t$ , where  $t$  is an integer parameter.

Substituting the form of  $z$  into the original equation, we get

$$\begin{aligned} 77x + 63y &= 5 - 33(1 + 7t), \\ 11x + 9y &= -4 - 33t. \end{aligned}$$

Now, we consider this (parametrized) equation modulo 11:

$$\begin{aligned} 9y &\equiv -4 - 33t \pmod{11}, \\ -2y &\equiv -4 \pmod{11}, \\ y &\equiv 2 \pmod{11}. \end{aligned}$$

Therefore, this congruence is satisfied by integers  $y = 2 + 11s$  for any  $s \in \mathbb{Z}$ . Now, it only remains to calculate  $x$ :

$$\begin{aligned} 11x &= -4 - 33t - 9(2 + 11s), \\ 11x &= -22 - 33t - 9 \cdot 11s, \\ x &= -2 - 3t - 9s. \end{aligned}$$

We have found out that the equation is satisfied if and only if  $(x, y, z)$  is in the set

$$\{(-2 - 3t - 9s, 2 + 11s, 1 + 7t); s, t \in \mathbb{Z}\}.$$

Particular solutions can be obtained by evaluating the triple at concrete values of  $t, s$ . For instance, setting  $t = s = 0$  gives the triple  $(-2, 2, 1)$ ; putting  $t = -4, s = 1$  leads to  $(1, 13, -27)$ .

Of course, the unknowns can be eliminated in any order – the result may seem “syntactically” different, but it must still describe the same set of solutions (that is given by a particular coset of an appropriate subgroup (in our case, it is the subgroup  $(2, 2, 1) + (3, 0, 7)\mathbb{Z} + (-9, 11, 0)\mathbb{Z}$ ) in the commutative group  $\mathbb{Z}^3$ , which is an apparent analog to the fact that the solution of such an equation over a field forms an affine subspace of the corresponding vector space).  $\square$

**Other types of Diophantine equations reducible to congruences.** Some Diophantine equations are such that one of the unknowns can be expressed explicitly as a function of the other ones. In this case, it makes sense to examine for which integer arguments it holds that the value of the function is also an integer.

For instance, having an equation of the form

$$mx_n = f(x_1, \dots, x_{n-1}),$$

where  $m$  is a natural number and  $f(x_1, \dots, x_{n-1}) \in \mathbb{Z}[x_1, \dots, x_{n-1}]$  is a polynomial with integer coefficients, an  $n$ -tuple of integers  $x_1, \dots, x_n$  is a solution of it if and only if

$$f(x_1, \dots, x_{n-1}) \equiv 0 \pmod{m}.$$

**11.E.2.** Solve the Diophantine equation  $x(x + 3) = 4y - 1$ .

**Solution.** The equation can be rewritten as  $4y = x^2 + 3x + 1$ . Now, we will solve the congruence

$$x^2 + 3x + 1 \equiv 0 \pmod{4}.$$

This congruence has no solution since for any integer  $x$ , the polynomial  $x^2 + 3x + 1$  evaluates to an odd integer (the

**11.4.14. Lemma (Gauss).** If  $p$  is an odd prime,  $a \in \mathbb{Z}$ ,  $p \nmid a$ , then the value of the Legendre symbol satisfies

$$\left(\frac{a}{p}\right) = (-1)^{\mu_p(a)}.$$

**PROOF.** For each integer  $i \in \{1, 2, \dots, \frac{p-1}{2}\}$ , we set a value  $m_i \in \{1, 2, \dots, \frac{p-1}{2}\}$  so that  $i \cdot a \equiv \pm m_i \pmod{p}$ . We can easily see that if  $k, l \in \{1, 2, \dots, \frac{p-1}{2}\}$  are different, then the values  $m_k, m_l$  are also different (the equality  $m_k = m_l$  would imply that  $k \cdot a \equiv \pm l \cdot a \pmod{p}$ , and hence  $k \equiv \pm l \pmod{p}$ , which cannot be satisfied unless  $k = l$ ).

Therefore, the sets  $\{1, 2, \dots, \frac{p-1}{2}\}$  and  $\{m_1, m_2, \dots, m_{\frac{p-1}{2}}\}$  coincide, which is also illustrated by the above example. Multiplying the congruences

$$\begin{aligned} 1 \cdot a &\equiv \pm m_1 \pmod{p}, \\ 2 \cdot a &\equiv \pm m_2 \pmod{p}, \end{aligned}$$

$\vdots$

$$\frac{p-1}{2} \cdot a \equiv \pm m_{\frac{p-1}{2}} \pmod{p}$$

leads to

$$\frac{p-1}{2}! \cdot a^{\frac{p-1}{2}} \equiv (-1)^\mu \cdot \frac{p-1}{2}! \pmod{p},$$

since there are exactly  $\mu$  negative values on the right-hand sides of the congruences. Dividing both sides by the integer  $\frac{p-1}{2}!$ , we get the wanted statement, making use of lemma 11.4.12, whence  $(a/p) \equiv a^{\frac{p-1}{2}} \pmod{p}$ .  $\square$

Now, with the help of Gauss’s lemma, we will prove the law of quadratic reciprocity.

**PROOF OF THE LAW OF QUADRATIC RECIPROCITY.** The first part has been already proven; for the rest, we first derive a lemma which will be utilized in the proof of both of the remaining parts.

Let  $a \in \mathbb{Z}$ ,  $p \nmid a$ ,  $k \in \mathbb{N}$  and let  $[x]$  and  $\langle x \rangle$  denote the integer part (i.e. floor) and the fractional part, respectively, of a real number  $x$ . Then,

$$\left[\frac{2ak}{p}\right] = \left[2\left[\frac{ak}{p}\right] + 2\left\langle\frac{ak}{p}\right\rangle\right] = 2\left[\frac{ak}{p}\right] + \left[2\left\langle\frac{ak}{p}\right\rangle\right].$$

This expression is odd if and only if  $\langle \frac{ak}{p} \rangle > \frac{1}{2}$ , which is if and only if the least residue (in absolute value) of the integer  $ak$  modulo  $p$  is negative (a watchful reader should notice the return from the calculations of (ostensibly) irrelevant expressions back to the conditions close to the Legendre symbol). The integer  $\mu_p(a)$  thus has the same parity (is congruent to, modulo 2) as  $\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{2ak}{p}\right]$ , whence (thanks to Gauss’s lemma) we get that

$$\left(\frac{a}{p}\right) = (-1)^{\mu_p(a)} = (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{2ak}{p}\right]}.$$

fact that the congruence is not solvable can also be established by trying out all four possible remainders modulo 4 into it).  
□

**11.E.3.** Solve the following equation in integers:



$$379x + 314y + 183y^2 = 210$$

**Solution.** The equation is linear in  $x$ , so the other unknown,  $y$ , must satisfy the congruence

$$183y^2 + 314y - 210 \equiv 0 \pmod{379}.$$

Now, we can complete the left-hand polynomial to square in order to get rid of the linear term. First of all, we must find a  $t \in \mathbb{Z}$  such that  $183 \cdot t \equiv 1 \pmod{379}$ . (In other words, we need to determine the inverse of the integer 183 modulo 379). For this purpose, we will use the Euclidean algorithm:

$$\begin{aligned} 379 &= 2 \cdot 183 + 13, \\ 183 &= 14 \cdot 13 + 1, \end{aligned}$$

whence

$$\begin{aligned} 1 &= 183 - 14 \cdot 13 = 183 - 14 \cdot (379 - 2 \cdot 183) = \\ &= 29 \cdot 183 - 14 \cdot 379. \end{aligned}$$

Therefore, we can take, for instance, the integer 29 to be our  $t$ . Now, multiplying both sides of the congruence by  $t = 29$  and rearranging it, we get an equivalent congruence:

$$y^2 + 10y - 26 \equiv 0 \pmod{379}$$

Now, we can complete the left-hand polynomial to square, which leads to (substituting  $z = y + 5$ )

$$\begin{aligned} (y + 5)^2 - 5^2 - 26 &\equiv 0 \pmod{379}, \\ z^2 &\equiv 51 \pmod{379}. \end{aligned}$$

Invoking the law of quadratic reciprocity, we calculate the Legendre symbol  $(51/379)$ :

$$\begin{aligned} \left(\frac{51}{379}\right) &= \left(\frac{3}{379}\right) \cdot \left(\frac{17}{379}\right) = \left(\frac{379}{3}\right) \cdot (-1) \cdot \left(\frac{379}{17}\right) \cdot (+1) = \\ &= \left(\frac{1}{3}\right) \cdot (-1) \cdot \left(\frac{5}{17}\right) = (1) \cdot (-1) \cdot \left(\frac{17}{5}\right) \cdot (+1) = \\ &= (-1) \cdot \left(\frac{2}{5}\right) = (-1) \cdot (-1) = 1, \end{aligned}$$

whence it follows that the congruence is solvable, and, in particular, it has two solutions modulo 379.

The proposition of exercise 11.D.38 implies that the solutions are of the form

$$z \equiv \pm 51^{\frac{380}{4}},$$

where  $51^3 \equiv 1 \pmod{379}$ , whence  $51^{95} = (51^3)^{31} \cdot 51^2 \equiv -52 \pmod{379}$ . The solution is thus  $z \equiv \pm 52 \pmod{379}$ , which gives for the original unknown that

$$y \equiv 47 \pmod{379}, \quad y \equiv -57 \pmod{379}.$$

Therefore, the given Diophantine equation is satisfied by those pairs  $(x, y)$  with  $y \in \{47 + 379 \cdot k; k \in \mathbb{Z}\} \cup \{-57 +$

Furthermore, if  $a$  is odd, then  $a + p$  is even and we get

$$\begin{aligned} \left(\frac{2a}{p}\right) &= \left(\frac{2a + 2p}{p}\right) = \left(\frac{4 \frac{a+p}{2}}{p}\right) = \left(\frac{2}{p}\right)^2 \cdot \left(\frac{a+p}{p}\right) \\ &= (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{(a+p)k}{p}\right]} = (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{ak}{p}\right]} \cdot (-1)^{\sum_{k=1}^{\frac{p-1}{2}} k}. \end{aligned}$$

Since the sum of the arithmetic series  $\sum_{k=1}^{\frac{p-1}{2}} k$  is  $\frac{1}{2} \frac{p-1}{2} \frac{p+1}{2} = \frac{p^2-1}{8}$ , we get (for  $a$  odd) the relation

$$\left(\frac{2}{p}\right) \cdot \left(\frac{a}{p}\right) = (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{ak}{p}\right]} \cdot (-1)^{\frac{p^2-1}{8}},$$

which, for  $a = 1$ , gives the wanted statement of item 2.

By part (2), which we have already proved, and the previous equality, we now get for odd integers  $a$  that

$$(1) \quad \left(\frac{a}{p}\right) = (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{ak}{p}\right]}.$$

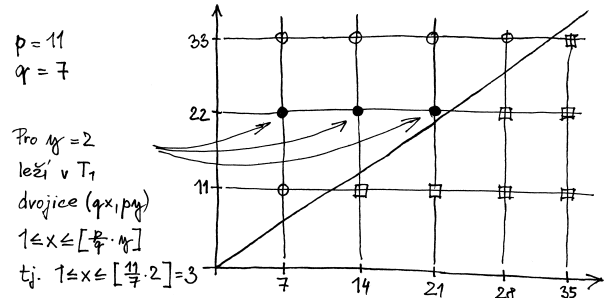
Now, let us consider, for given primes  $p \neq q$ , the set

$$\begin{aligned} T &= \{q \cdot x; x \in \mathbb{Z}, 1 \leq x \leq (p-1)/2\} \times \\ &\quad \times \{p \cdot y; y \in \mathbb{Z}, 1 \leq y \leq (q-1)/2\}. \end{aligned}$$

We apparently have  $|T| = \frac{p-1}{2} \cdot \frac{q-1}{2}$ . We will show that we also have

$$(-1)^{|T|} = (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{pk}{q}\right]} \cdot (-1)^{\sum_{k=1}^{\frac{p-1}{2}} \left[\frac{qk}{p}\right]},$$

which will be sufficient thanks to the above.



Since the equality  $qx = py$  can happen for no pair of  $x, y$  from the permissible domain, the set  $T$  can be partitioned to (disjoint) subsets  $T_1$  and  $T_2$  so that  $T_1 = T \cap \{(u, v); u, v \in \mathbb{Z}, u < v\}$ ,  $T_2 = T \setminus T_1$ . Clearly,  $|T_1|$  is the number of pairs  $(qx, py)$  for which  $x < \frac{p}{q}y$ . Since  $\frac{p}{q}y \leq \frac{p}{q} \cdot \frac{q-1}{2} < \frac{p}{2}$ , we have  $\left[\frac{p}{q}y\right] \leq \frac{p-1}{2}$ . For a fixed  $y$ , in  $T_1$ , there are thus exactly those pairs  $(qx, py)$  for which  $1 \leq x \leq \left[\frac{p}{q}y\right]$ ; hence  $|T_1| = \sum_{y=1}^{(q-1)/2} \left[\frac{p}{q}y\right]$ . Analogously,  $|T_2| = \sum_{x=1}^{(p-1)/2} \left[\frac{q}{p}x\right]$ .

By (1), we thus have  $\left(\frac{p}{q}\right) = (-1)^{|T_1|}$  and  $\left(\frac{q}{p}\right) = (-1)^{|T_2|}$ , which finishes the proof of the law of quadratic reciprocity. □

$379 \cdot k; k \in \mathbb{Z}$  and  $x = \frac{1}{379} \cdot (210 - 314y - 183y^2)$ ; e. g.  $(-1105, 47)$  or  $(-1521, -57)$  (which are the only solutions with  $|x| < 10^5$ ).  $\square$

**11.E.4.** Solve the equation  $2^x = 1 + 3^y$  in integers.

**Solution.** If  $y < 0$ , then  $1 < 1 + 3^y < 2$ , whence  $0 < x < 1$ , so  $x$  could not be an integer. Therefore,  $y \geq 0$ , hence  $2^x = 1 + 3^y \geq 2$  and  $x \geq 1$ . We will show that we also must have  $x \leq 2$ . If not (i.e., if  $x \geq 3$ ), then we would have

$$1 + 3^y = 2^x \equiv 0 \pmod{8},$$

whence it follows that

$$3^y \equiv -1 \pmod{8}.$$

However, this is impossible since the order of 3 modulo 8 equals 2, so the powers of three are congruent to 3 and 1 only. Now, it remains to examine the possibilities  $x = 1$  and  $x = 2$ .

For  $x = 1$ , we get

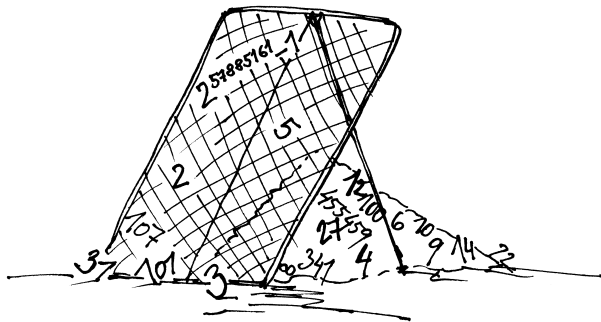
$$3^y = 2^1 - 1 = 1,$$

hence  $y = 0$ . If  $x = 2$ , we have

$$3^y = 2^2 - 1 = 3,$$

whence  $y = 1$ . Thus, the equations has two solutions:  $x = 1, y = 0$ ; and  $x = 2, y = 1$ .  $\square$

**F. Primality tests**



**11.F.1. Mersenne primes.** The following problems are in deep connection with testing Mersenne numbers for primality.



For any  $q \in \mathbb{N}$ , consider the integer  $M_q = 2^q - 1$  and prove:

- i) If  $q$  is composite, then so is  $M_q$ .
- ii) If  $q$  is a prime,  $q \equiv 3 \pmod{4}$ , then  $2q + 1$  divides  $M_q$  if and only if  $2q + 1$  is a prime (hence it follows that if  $q \equiv 3 \pmod{4}$  is a Sophie Germain prime<sup>3</sup>, then  $M_q$  is not a prime).
- iii) If a prime  $p$  divides  $M_q$ , then  $p \equiv \pm 1 \pmod{8}$  and  $p \equiv 1 \pmod{q}$ .

**Solution.**

<sup>3</sup>See Wikipedia, *Sophie Germain prime*, [http://en.wikipedia.org/wiki/Sophie\\_Germain\\_prime](http://en.wikipedia.org/wiki/Sophie_Germain_prime) (as of July 28, 2013, 14:43 GMT).



The evaluation of the Legendre symbol (as we saw in the example above) allows us to only use the law of quadratic reciprocity for primes, so it forces us to factor integers to primes, which is a very hard operation from the computational point of view. This can be mended by extending the definition of the Legendre symbol to the so-called *Jacobi symbol* with similar properties.

**Definition.** Let  $a \in \mathbb{Z}, b \in \mathbb{N}, 2 \nmid b$ . Let  $b$  factor as  $b = p_1 p_2 \cdots p_k$  to (odd) primes (here, we exceptionally do not group the same primes to a power of the prime, rather we write each one explicitly, e.g.  $135 = 3 \cdot 3 \cdot 3 \cdot 5$ ). The symbol

$$\left(\frac{a}{b}\right) = \left(\frac{a}{p_1}\right) \cdot \left(\frac{a}{p_2}\right) \cdots \left(\frac{a}{p_k}\right)$$

is called the *Jacobi symbol*.

We show in the practical column that the Jacobi symbol has similar properties as the Legendre one. However, there is a substantial aberration – it is not generally true that  $(a/b) = 1$  implies that the congruence  $x^2 \equiv a \pmod{b}$  is solvable.

**Example.**

$$\left(\frac{2}{15}\right) = \left(\frac{2}{3}\right) \cdot \left(\frac{2}{5}\right) = (-1) \cdot (-1) = 1,$$

but the congruence

$$x^2 \equiv 2 \pmod{15}$$

has no solution (the congruence  $x^2 \equiv 2$  is solvable neither modulo 3 nor modulo 5).

**Theorem** (Law of quadratic reciprocity for the Jacobi symbol). *Let  $a, b \in \mathbb{N}$  be odd integers. Then,*

- (1)  $\left(\frac{-1}{a}\right) = (-1)^{\frac{a-1}{2}}$ ,
- (2)  $\left(\frac{2}{a}\right) = (-1)^{\frac{a^2-1}{8}}$ ,
- (3)  $\left(\frac{a}{b}\right) = \left(\frac{b}{a}\right) \cdot (-1)^{\frac{a-1}{2} \cdot \frac{b-1}{2}}$ .

**PROOF.** The proof is simple, utilizing the law of quadratic reciprocity for the Legendre symbol. See exercise 11.D.35.  $\square$

There is another application of the law of quadratic reciprocity in a certain sense – we can consider the question: *For which primes is a given integer  $a$  a quadratic residue?* We are already able to answer this question for  $a = 2$ , for example. The first step in answering this question is to do so for primes since the answer for composite values of  $a$  depends on the factorization of the integer  $a$ .

**Theorem.** *Let  $q$  be an odd prime.*

- If  $q \equiv 1 \pmod{4}$ , then  $q$  is a quadratic residue modulo those primes  $p$  which satisfy  $p \equiv r \pmod{q}$ , where  $r$  is a quadratic residue modulo  $q$ .
- If  $q \equiv 3 \pmod{4}$ , then  $q$  is a quadratic residue modulo those primes  $p$  which satisfy  $p \equiv \pm b^2 \pmod{4q}$ , where  $b$  is odd and coprime to  $q$ .

- i) If  $n \mid q$ , then it follows from exercise 11.A.7 that  $2^n - 1 \mid 2^q - 1$ , so  $M_n \mid M_q$ . Therefore,  $M_q$  is not a prime for  $n > 1$ .
- ii) Let  $n = 2q+1$  be a divisor of  $M_q$ . We will show that  $n$  is a prime invoking Lucas' theorem 11.6.10. Since  $n - 1 = 2q$  has only two prime divisors, it suffices to find compositeness witnesses for the integers 2 and  $q$ . We have  $2^{\frac{n-1}{q}} = 2^2 \not\equiv 1 \pmod{n}$ ,  $(-2)^{\frac{n-1}{2}} = -2^q \equiv -1 \not\equiv 1 \pmod{n}$ , thanks to the assumption  $n \mid M_q = 2^q - 1$ . Further, since  $(-2)^{n-1} = 2^{n-1} = 2^{2q} - 1 = (2^q + 1)M_q \equiv 0 \pmod{n}$ , it follows from Lucas' theorem that  $n$  is a prime.

Now, let  $p = 2q + 1 \equiv -1 \pmod{8}$  be a prime. Since  $(2/p) = 1$ , there exists an  $m$  such that  $2 \equiv m^2 \pmod{p}$ . Hence,  $2^q \equiv 2^{\frac{p-1}{2}} \equiv m^{p-1} \equiv 1 \pmod{p}$ , so  $p \mid 2^q - 1 = M_q$ .

- iii) If  $p \mid M_q = 2^q - 1$ , then the order of 2 modulo  $p$  must divide the prime  $q$ , hence it equals  $q$ . Therefore,  $q \mid p-1$ , and there exists a  $k \in \mathbb{Z}$  such that  $2qk = p - 1$ . Altogether, we get

$$(2/p) \equiv 2^{\frac{p-1}{2}} \equiv 2^{qk} \equiv 1 \pmod{p},$$

i.e.,  $p \equiv \pm 1 \pmod{8}$ . □

**11.F.2.** For each of the following Mersenne numbers, determine whether it is prime or composite:  $2^{11} - 1$ ,  $2^{15} - 1$ ,  $2^{23} - 1$ ,  $2^{29} - 1$ , and  $2^{83} - 1$ .



**Solution.** In the case of the integer  $2^{15} - 1$ , the exponent is composite; therefore, the whole integer is composite as well (we even know that it is divisible by  $2^3 - 1$  and  $2^5 - 1$ ). In the other cases, the exponent is always a prime. We can notice that these primes, namely  $q = 11, 23, 29$ , and  $83$ , are even Sophie Germain primes (i.e.,  $2q + 1$  is also a prime). It thus follows from part (ii) of the previous exercise that  $23 \mid 2^{11} - 1$ ,  $47 \mid 2^{23} - 1$ , and  $167 \mid 2^{83} - 1$ .

We cannot use this proposition for the last case since  $29 \not\equiv 3 \pmod{4}$  and, indeed,  $59 \nmid 2^{29} - 1$ . Now, however, it follows from part (iii) of the above exercise that if there is a prime  $p$  which divides  $2^{89} - 1$ , then it must satisfy

$$\begin{aligned} p &\equiv \pm 1 \pmod{8} \\ p &\equiv 1 \pmod{29}, \end{aligned}$$

i.e.,  $p \equiv 1 \pmod{232}$  or  $p \equiv 175 \pmod{232}$ . If we are looking for a prime divisor of the integer  $n = 2^{29} - 1 = 536\,870\,911$ , then it suffices to check the primes (of the above form) up to  $\sqrt{n} \approx 23\,170$ . There are 50 of them, so we are able to decide whether  $n$  is a prime quite easily (even with paper and pencil). In this case, fortunately,  $n$  is divisible already by the least prime, 233. □

**11.F.3.** Show that the integer 341 is a Fermat pseudoprime to base 2, yet it is not a Euler-Jacobi pseudoprime to base 2. Further, prove that the integer 561 is a Euler-Jacobi pseudoprime to base 2, but not to base 3. Prove that, on the other hand, the integer



**PROOF.** The first theorem follows trivially from the law of quadratic reciprocity. Let us consider  $q \equiv 3 \pmod{4}$ , i.e.,  $(q/p) = (-1)^{\frac{p-1}{2}}(p/q)$ . First of all, let  $p \equiv +b^2 \pmod{4q}$ , where  $b$  is odd, and hence  $b^2 \equiv 1 \pmod{4}$ . Then,  $p \equiv b^2 \equiv 1 \pmod{4}$  and  $p \equiv b^2 \pmod{q}$ . Therefore,  $(-1)^{\frac{p-1}{2}} = 1$  and  $(p/q) = 1$ , whence  $(q/p) = 1$ . Now, if  $p \equiv -b^2 \pmod{4q}$ , then we similarly get that  $p \equiv -b^2 \equiv 3 \pmod{4}$  and  $p \equiv -b^2 \pmod{q}$ . Therefore,  $(-1)^{\frac{p-1}{2}} = -1$  and  $(p/q) = -1$ , whence we get again that  $(q/p) = 1$ .

For the opposite way, suppose that  $(q/p) = 1$ . There are two possibilities – either  $(-1)^{\frac{p-1}{2}} = 1$  and  $(p/q) = 1$ , or  $(-1)^{\frac{p-1}{2}} = -1$  and  $(p/q) = -1$ . In the former case, we have  $p \equiv 1 \pmod{4}$  and there is a  $b$  such that  $p \equiv b^2 \pmod{q}$ . Further, we can assume without loss of generality that  $b$  is odd (if not, we could have taken  $b + q$  instead). However, then we get  $b^2 \equiv 1 \equiv p \pmod{4}$ , and altogether  $p \equiv b^2 \pmod{4q}$ .

In the latter case, we have  $p \equiv 3 \pmod{4}$  and  $(-p/q) = (-1/q)(p/q) = (-1)(-1) = 1$ . Therefore, there is a  $b$  (which can also be chosen so that it is odd) such that  $-p \equiv b^2 \pmod{q}$ . We thus get  $-b^2 \equiv 3 \equiv p \pmod{4}$ , and altogether  $p \equiv -b^2 \pmod{4q}$ . □

### 5. Diophantine equations

It is as early as in the third century AD when Diophantus of Alexandria dealt with miscellaneous equations while admitting only integers as solutions. And there is no wonder – in many practical problems that lead to equations, non-integer solutions may fail to have a meaningful interpretation. As an example, we can consider the problem of how to pay an exact amount of money with coins of given values.

In honor of Diophantus, equations for which we are interested in integer solutions only are called *Diophantine equations*.

Another nice example of a Diophantine equation is Euler's relation

$$v - e + f = 2$$

from graph theory, connecting the number of vertices, edges, and faces of a planar graph. Furthermore, if we restrict ourselves to regular graphs only, we get to the problem about existence of the so-called Platonic solids, which can be smartly described just as a solution of this Diophantine equation – for more information, see 13.1.22.

Unfortunately, there is no universal method for solving this kind of equations. There is even no method (algorithm) to decide whether a given polynomial Diophantine equation has a solution. This question is well-known as *Hilbert's tenth problem*, and the proof of algorithmic unsolvability of this problem was given by Юрий Матиясевич (Yuri Matiyasevich) in 1970.<sup>7</sup>

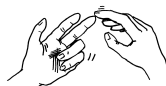
However, there are cases in which we are able to find the solution of a Diophantine equation, or – at least – to reduce

<sup>7</sup>See the elementary text M. Davis, *Hilbert's Tenth Problem is Unsolv-able*, The American Mathematical Monthly 80(3): 233–269. 1973.

121 is a Euler-Jacobi pseudoprime to base 3, but not to base 2.

**Solution.** The integer 341 is a Fermat pseudoprime to base 2 since  $2^{10} \equiv 1 \Rightarrow 2^{340} \equiv 1 \pmod{341}$ . It is not a Euler-Jacobi pseudoprime since  $2^{170} \equiv 1 \pmod{341}$ , but  $(\frac{2}{341}) = -1$ , which follows from the fact that  $341 \equiv -3 \pmod{8}$ . For the integer 561, we have  $2^{280} \equiv 1 \pmod{561}$  and  $(\frac{2}{561}) = 1$ , since  $561 \equiv 1 \pmod{8}$ . Therefore, it is a Euler-Jacobi pseudoprime to base 2. But not to base 3, since  $3 \mid 561$ . On the other hand, the integer 121 satisfies  $3^5 \equiv 1 \pmod{121} \Rightarrow 3^{60} \equiv 1 \pmod{121}$  and  $(\frac{3}{121}) = 1$ , but  $2^{60} \equiv 89 \not\equiv 1 \pmod{121}$ .  $\square$

**11.F.4.** Prove that the integers 2465, 2821, and 6601 are Carmichael numbers, i.e., denoting any of them as  $n$ , then every integer  $a$  coprime to  $n$  satisfies



$$a^{n-1} \equiv 1 \pmod{n}.$$

**Solution.** We have  $2465 = 5 \cdot 17 \cdot 29$ ,  $2821 = 7 \cdot 13 \cdot 31$ ,  $6601 = 7 \cdot 23 \cdot 41$ , and the proposition follows from Korselt's criterion 11.6.6 since all of the integers 4, 16, 28 divide  $2464 = 2^5 \cdot 7 \cdot 11$ , all of the integers 6, 12, 30 divide  $2820 = 2^2 \cdot 3 \cdot 5 \cdot 47$ , and 6, 22, 40 divide  $6600 = 2^3 \cdot 3 \cdot 5^2 \cdot 11$ .  $\square$

**11.F.5.** Prove that the integer 2047 is a strong pseudoprime to base 2, but not to base 3. Further, prove that the integer 1905 is a Euler-Jacobi pseudoprime to base 2 but not a strong pseudoprime to this base.



**Solution.** In order to verify whether 2047 is a strong pseudoprime to base 2, we factor

$$(2^{2046} - 1) = (2^{1023} - 1)(2^{1023} + 1).$$

Since  $2^{1023} \equiv 1 \pmod{2047}$ , the statement is true. However, it is not a strong pseudoprime to base 3 as

$$3^{1023} \equiv 1565 \not\equiv \pm 1 \pmod{2047}.$$

Notice that for the integer 2047, the strong pseudoprimality test is identical to the Euler one (this is because the integer 2046 is not divisible by four).

The integer 1905 is a Euler-Jacobi pseudoprime to base 2 since  $2^{1904/2} \equiv 1 \pmod{1905}$  and the Jacobi symbol  $(2/1905)$  is equal to 1. Since  $1904 = 2^4 \cdot 7 \cdot 17$ , 1905 will be a strong pseudoprime to base 2 only if at least one of the following congruences holds:

$$\begin{aligned} 2^{952} &\equiv -1 \pmod{1905}, \\ 2^{476} &\equiv -1 \pmod{1905}, \\ 2^{238} &\equiv -1 \pmod{1905}, \\ 2^{119} &\equiv \pm 1 \pmod{1905}. \end{aligned}$$

However,  $2^{952} \equiv 2^{476} \equiv 1 \pmod{1905}$ ,  $2^{238} \equiv 1144 \pmod{1905}$ , and  $2^{119} \equiv 128 \pmod{1905}$ . Therefore, 1905 is not a strong pseudoprime to base 2.  $\square$

the problem to solving congruences, which is besides the already mentioned applications another motivation for studying them. Now, we will describe several such types of Diophantine equations.

**Linear Diophantine equation.** A linear Diophantine equation is an equation of the form



$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b,$$

where  $x_1, \dots, x_n$  are unknowns and  $a_1, \dots, a_n, b$  are given non-zero integers.



*We can see that the ability to solve Diophantine equations is sometimes important in "practical" life as well, as is proved by Bruce Willis and Samuel Jackson in Die Hard with a Vengeance, where they have to do away with a bomb using 4 gallons of water, having only 3- and 5-gallon containers at their disposal. A mathematician would say that the gentlemen were to find a solution of the Diophantine equation  $3x + 5y = 4$ .*

One can use congruences in order to solve these equations. Apparently, it is necessary for the equation to be solvable that the integer  $d = (a_1, \dots, a_n)$  divides  $b$ . Provided that, dividing both sides of the equation by the number  $d$  leads to an equivalent equation

$$a'_1x_1 + a'_2x_2 + \dots + a'_nx_n = b',$$

where  $a'_i = a_i/d$  for  $i = 1, \dots, n$  and  $b' = b/d$ . Here, we have

$$d \cdot (a'_1, \dots, a'_n) = (da'_1, \dots, da'_n) = (a_1, \dots, a_n) = d,$$

so  $(a'_1, \dots, a'_n) = 1$ .

Further, we will show that the equation

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b,$$

where  $a_1, a_2, \dots, a_n, b$  are integers such that  $(a_1, \dots, a_n) = 1$ , always have a solution in integers and all such solutions can be described in terms of  $n - 1$  integer parameters.

We will prove this proposition by mathematic induction on  $n$ , the number of unknowns. The situation is trivial for  $n = 1$  – there is a unique solution (which does not depend

**11.F.6.** Applying Pocklington-Lehmer test 11.6.11, show that 1321 is a prime.



**Solution.** Let us set  $N = 1321$ , then  $N - 1 = 1320 = 2^3 \cdot 3 \cdot 5 \cdot 11$ .

For the sake of simplicity, we will assume that the trial division is executed only for primes below 10, then  $F = 2^3 \cdot 3 \cdot 5 = 120$ ,  $U = 11$ , where  $(F, U) = (120, 11) = 1$ .

In order to prove the primality of 1321 by the Pocklington-Lehmer test, we need to find a primality witness  $a_p$  for each  $p \in \{2, 3, 5\}$ .

Since  $(2^{\frac{1320}{3}} - 1, 1321) = 1$  and  $(2^{\frac{1320}{5}} - 1, 1321) = 1$ , we can lay  $a_3 = a_5 = 2$ . However, for  $p = 2$ , we have  $(2^{\frac{1320}{2}} - 1, 1321) = 1321$ , so we have to look for another primality witness. We can take  $a_2 = 7$  since  $(7^{\frac{1320}{2}} - 1, 1321) = 1$ . In both cases, we have  $2^{1320} \equiv 7^{1320} \equiv 1 \pmod{1321}$ . The primality witnesses of the integer 1321 are thus  $a_2 = 7, a_3 = a_5 = 2$ . Instead, we could also have chosen for all primes  $p$  the same number (e. g. 13), which is a primitive root modulo 1321.  $\square$

**11.F.7.** Factor the integer 221 to primes by Pollard's  $\rho$ -method. Use the function  $f(x) = x^2 + 1$  with initial value  $x_0 = 2$ .



**Solution.** Let us set  $x = y = 2$ . The procedure from 11.6.14 gives:

$x := f(x)$	$y := f(f(y))$	$( x - y , 221) \pmod{221}$
5	26	1
26	197	1
14	104	1
197	145	13

We have thus found a non-trivial divisor, so now it is easy to calculate  $221 = 13 \cdot 17$ .  $\square$

**11.F.8.** Find a non-trivial divisor of the integer 455459.

**Solution.** Consider the function  $f(x) = x^2 + 1$  (we silently assume that this function behaves randomly modulo an unknown prime divisor  $p$  of the integer  $n$  and has the required properties). In the particular iterations, we compute  $a \leftarrow f(a) \pmod{n}$ ,  $b \leftarrow f(f(b)) \pmod{n}$  while evaluating  $d = (a - b, n)$ .

a	b	d
5	26	1
26	2871	1
677	179685	1
2871	155260	1
44380	416250	1
179685	43670	1
121634	164403	1
155260	247944	1
44567	68343	743

We have found a divisor 743, and now we can easily compute that  $455459 = 613 \cdot 743$ .  $\square$

on any parameters). Further, let  $n \geq 2$  and suppose that the statement holds for equations having  $n - 1$  unknowns. Denoting  $d = (a_1, \dots, a_{n-1})$ , any  $n$ -tuple  $x_1, \dots, x_n$  that satisfies the equation must also satisfy the congruence

$$a_1x_1 + a_2x_2 + \dots + a_nx_n \equiv b \pmod{d}.$$

Since  $d$  is the greatest common divisor of the integers  $a_1, \dots, a_{n-1}$ , this congruence is of the form

$$a_nx_n \equiv b \pmod{d},$$

which (since  $(d, a_n) = (a_1, \dots, a_n) = 1$ ) has a unique solution

$$x_n \equiv c \pmod{d},$$

where  $c$  is a suitable integer, i.e.,  $x_n = c + dt$ , where  $t \in \mathbb{Z}$  is arbitrary. Substituting into the original equation and refining it leads to the equation

$$a_1x_1 + \dots + a_{n-1}x_{n-1} = b - a_nc - a_ndt$$

with  $n - 1$  unknowns and one parameter,  $t$ . However, the number  $(b - a_nc)/d$  is an integer, so we can divide the equation by  $d$ . This leads to

$$a'_1x_1 + \dots + a'_{n-1}x_{n-1} = b',$$

where  $a'_i = a_i/d$  for  $i = 1, \dots, n - 1$  and  $b' = ((b - a_nc)/d) - a_nt$ , satisfying

$$(a'_1, \dots, a'_{n-1}) = (da'_1, \dots, da'_{n-1}) \cdot \frac{1}{d} = (a_1, \dots, a_{n-1}) \cdot \frac{1}{d} = 1.$$

By the induction hypothesis, this equation has, for any  $t \in \mathbb{Z}$ , a solution which can be described in terms of  $n - 2$  integer parameters (different from  $t$ ), which together with the condition  $x_n = c + dt$  gives what we wanted.

**11.5.1. Pythagorean equation.** In this section, we will deal with enumeration of all right triangles with integer side lengths. This is a Diophantine equation where we will only seldom use the methods described above; nevertheless, we will look at it in detail.



The task is to solve the equation

$$x^2 + y^2 = z^2$$

in integers.

**Solution.** Clearly, we can assume that  $(x, y, z) = 1$  (otherwise, we simply divide both sides of the equation by the integer  $d = (x, y, z)$ ).

Further, we can show that the integers  $x, y, z$  are pairwise coprime: if there were a prime  $p$  dividing two of them, then we can easily see that it would have to divide the third one as well, which it may not according to our assumption. Therefore, at most one of the integers  $x, y$  is even. If neither of them were, we would get

$$z^2 \equiv x^2 + y^2 \equiv 1 + 1 \pmod{8},$$

which is impossible (see exercise 11.A.2). Altogether, we get that exactly one of the integers  $x, y$  is even. However, since the roles of these integers in the equation are symmetric, we

**G. Encryption**

**11.G.1. RSA.** We have overheard that the integers 29, 7, 21 were sent by means of RSA with public key (7, 33). Try to break the cipher and find the messages (integers) that were originally sent.



**Solution.** In order to find the private key  $d$ , we need to solve the congruence  $7d \equiv 1 \pmod{\varphi(33)}$ . However, since the integer 33 is quite small, we can factor it and easily compute that  $\varphi(33) = (3 - 1)(11 - 1) = 20$ . We are thus looking for a  $d$  such that  $7d \equiv 1 \pmod{20}$ , which is satisfied by  $d \equiv 3 \pmod{20}$ . Since  $29^3 \equiv (-4)^3 \equiv 2$ ,  $7^3 \equiv 13$ , and  $21^3 \equiv 21 \pmod{33}$ , the messages that were encrypted are 2, 13, and 21.  $\square$

**Attacks against RSA.**



Using so-called Fermat's factorization method, we can try to factor  $n = p \cdot q$  if we think that the difference between  $p$  and  $q$  is small.

Then,

$$n = \left(\frac{p+q}{2}\right)^2 - \left(\frac{p-q}{2}\right)^2,$$

where  $s = (p - q)/2$  is small and  $t = (p + q)/2$  is only a bit greater than  $\sqrt{n}$ . Therefore, it suffices to check whether<sup>4</sup>  $t = \lceil\sqrt{n}\rceil, t = \lceil\sqrt{n}\rceil + 1, t = \lceil\sqrt{n}\rceil + 2, \dots$ , until  $t^2 - n$  is a square (this condition can, of course, be checked efficiently).

**11.G.2.** Now, we will try to factor the integer  $n = 23104222007$  this way. (We anticipate that it is a product of two close primes.)



**Solution.** We compute

$$\sqrt{n} \approx 152000,731$$

and check the candidates for  $t$ :

For  $t = 152001$ , we have  $\sqrt{t^2 - n} \approx 286,345$ .

For  $t = 152002$ , it is  $\sqrt{t^2 - n} \approx 621,287$ .

For  $t = 152003$ ,  $\sqrt{t^2 - n} \approx 830,664$ .

Finally, for  $t = 152004$ , we get  $\sqrt{t^2 - n} = 997 \in \mathbb{Z}$ .

Therefore,  $s = 997$  and we can easily calculate the prime divisors of  $n$ :  $p = t + s = 153001, q = t - s = 151007$ .  $\square$

**11.G.3.** The RSA modulus  $n = p \cdot q$  can also be easily factored if the integer  $\varphi(n)$  is known (compromised). Then,

$$\varphi(n) = (p-1)(q-1) = pq - (p+q) + 1, \text{ odkud } p+q = n+1 - \varphi(n).$$

We are thus to find two integers whose sum and product are known, which can be done by Viète's formulas relating the roots and the coefficients of a polynomial, whence it follows that  $p$  and  $q$  are the roots of the polynomial

$$x^2 - (n + 1 - \varphi(n))x + n.$$

<sup>4</sup>The symbol  $\lceil x \rceil$  denotes the *ceiling* if a real number  $x$ , i.e., the integer which satisfies  $\lceil x \rceil - 1 < x \leq \lceil x \rceil$ .

can, without loss of generality, select  $x$  to be even and set  $x = 2r, r \in \mathbb{N}$ . Hence, we have

$$4r^2 = z^2 - y^2,$$

so

$$r^2 = \frac{z+y}{2} \cdot \frac{z-y}{2}.$$

Now, let us denote  $u = \frac{1}{2}(z + y), v = \frac{1}{2}(z - y)$  (then, the inverse substitution is  $z = u + v, y = u - v$ ). Since  $y$  is coprime to  $z$ , so is  $u$  to  $v$  (if there were a prime  $p$  dividing both  $u$  and  $v$ , then it would divide their sum as well as their difference, i.e., the integers  $y$  and  $z$ ). It follows from

$$r^2 = u \cdot v$$

that there are coprime positive integers  $a, b$  such that  $u = a^2, v = b^2$ . Moreover, since  $u > v$ , we must have  $a > b$ . Altogether, we get

$$x = 2r = 2ab,$$

$$y = u - v = (a^2 - b^2),$$

$$z = u + v = (a^2 + b^2),$$

which indeed satisfies the given equation for any coprime  $a, b \in \mathbb{N}$  with  $a > b$ . Further solutions can be obtained by interchanging  $x$  and  $y$ . Finally, relinquishing the condition  $(x, y, z) = 1$ , each solution will yield infinitely many more if we multiply each of its component by a fixed positive integer  $d$ .  $\square$

**11.5.2. Fermat's Last Theorem for  $n = 4$ .** Thanks to the parametrization of Pythagorean triples, we will be able to prove that the famous Fermat's Last Theorem



$$x^n + y^n = z^n$$

has no solution for  $n = 4$  in integers. For this task it is sufficient to prove that the equation  $x^4 + y^4 = z^2$  has no solution in  $\mathbb{N}$ .

**Solution.** We will use the so-called method of infinite descent, which was introduced by Pierre de Fermat. This method utilizes the fact that every non-empty set of natural numbers has a least element (in other words,  $\mathbb{N}$  is a well-ordered set).

Therefore, suppose that the set of solutions of the equation  $x^4 + y^4 = z^2$  is non-empty and let  $(x, y, z)$  denote (any) solution with  $z$  as small as possible. The integers  $x, y, z$  are thus pairwise distinct. Since the equation can be written in the form

$$(x^2)^2 + (y^2)^2 = z^2,$$

it follows from the previous exercise that there exist  $r, s \in \mathbb{N}$  such that

$$x^2 = 2rs, \quad y^2 = r^2 - s^2, \quad z = r^2 + s^2.$$

Hence,  $y^2 + s^2 = r^2$ , where  $(y, s) = 1$  (if there were a prime  $p$  dividing both  $y$  and  $s$ , then it would divide  $x$  as well as  $z$ , which contradicts that they are coprime). Making the



**11.G.4.** Consider (as above) the integer  $n = 23104222007$  and factor it with the additional knowledge that  $\varphi(n) = 23103918000$ .



**Solution.** Following the procedure described above, we get the quadratic equation

$$x^2 - 304008x + 23104222007 = 0,$$

whose solutions are

$$p = \frac{1}{2}(304008 + \sqrt{304008^2 - 4 \cdot 23104222007}) = 153001,$$

$$q = \frac{1}{2}(304008 - \sqrt{304008^2 - 4 \cdot 23104222007}) = 151007. \quad \square$$

**11.G.5. ElGamal.** Martin and John want to communicate using the ElGamal encryption (designed by Egyptian mathematician Taher ElGamal, who was inspired by the Diffie-Hellman key exchange protocol). Martin chose the prime 41 and its primitive root  $g = 11$  as well as the integer 10. Then, he published the triple  $(41, 11, A)$ , where  $A \equiv 11^{10} \pmod{41}$ ; he kept the integer 10 to himself – it is his private key. John used a public channel to send the pair  $(22, 6)$  to him. What is the original message John sent?



**Solution.** For completeness, we will first compute the whole public key  $A = 9$  (however, this integer was only needed by John when he encrypted the message for Martin; it is no longer necessary for decryption). The message  $M$  can be obtained as  $M \equiv (6/22^{10}) \pmod{41}$ . First, we compute

$$\begin{aligned} 22^{10} &\equiv 22^2 \cdot (22^2)^2 \cdot ((22^2)^2)^2 \equiv \\ &\equiv (-8) \cdot (-8)^2 \cdot (-8)^2 \equiv \\ &\equiv (-8) \cdot 23 \cdot 23 \equiv -9 \pmod{41} \end{aligned}$$

and  $(-9)^{-1} \equiv 9 \pmod{41}$ . Therefore, the decrypted message is the integer

$$M = 9 \cdot 6 \equiv 13 \pmod{41}. \quad \square$$

**11.G.6. Rabin cryptosystem.** Alice has chosen  $p = 23, q = 31$  as her private key in Rabin cryptosystem. The public key is  $n = pq = 713$ , then. Encrypt the message  $M = 327$  for Alice and show how Alice will decrypt it.



**Solution.** We compute  $C = (327)^2 \equiv 692 \pmod{713}$  and send this cipher to Alice. According to the decryption procedure, we determine

$$\begin{aligned} r &\equiv C^{(p+1)/4} \equiv 692^{\frac{23+1}{4}} \equiv 18 \pmod{23}, \\ s &\equiv C^{(q+1)/4} \equiv 692^{\frac{31+1}{4}} \equiv 14 \pmod{31}, \end{aligned}$$

and further the coefficients  $a, b$  into Bézout's identity  $23 \cdot a + 31 \cdot b = 1$  (using the Euclidean algorithm). We get  $a = -4, b = 3$ ; the candidates for the original message are thus

Pythagorean substitution once again, we get natural numbers  $a, b$  with ( $y$  is odd)

$$y = a^2 - b^2, \quad s = 2ab, \quad r = a^2 + b^2.$$

The inverse substitution leads to

$$x^2 = 2rs = 2 \cdot 2ab(a^2 + b^2),$$

and since  $x$  is even, we get

$$\left(\frac{x}{2}\right)^2 = ab(a^2 + b^2).$$

The integers  $a, b, a^2 + b^2$  are pairwise coprime (which can be derived easily from the fact that  $y$  is coprime to  $s$ ). Therefore, each of them is a square of a natural number:

$$a = c^2, \quad b = d^2, \quad a^2 + b^2 = e^2,$$

whence  $c^4 + d^4 = e^2$ , and since  $e \leq a^2 + b^2 = r < z$ , we get a contradiction with the minimality of  $z$ .  $\square$

## 6. Applications – calculation with large integers, cryptography

**11.6.1. Computational aspects of number theory.** In



many practical problems which utilize the results of number theory, it is necessary to execute one or more of the following computations fast:

- common arithmetic operations (sum, product, modulo) on integers;
- to determine the remainder of a (natural)  $n$ -th power of an integer  $a$  when divided by a given  $m$ ;
- to determine the multiplicative inverse of an integer  $a$  modulo  $m \in \mathbb{N}$ ;
- to determine the greatest common divisor of two integers (and the coefficients of corresponding Bézout's identity);
- to decide whether a given integer is a prime or composite number.
- to factor a given integer to primes.

Basic arithmetic operations are usually executed on large integers in the same way as we were taught at primary school, i.e., we add in *linear* time and multiply and divide with remainder in *quadratic* time. The multiplication, which is a base for many other operations, can be performed asymptotically more efficiently (there exist algorithms of the type *divide and conquer*) – for instance, the Karatsuba algorithm (1960), running in time  $\Theta(n^{\log_2 3})$  or the Schönhage-Strassen algorithm (1971), which runs in  $\Theta(n \log n \log \log n)$  and uses Fast Fourier Transforms – see also 7.2.5. Although it is asymptotically much better, in practice, it becomes advantageous for integers of at least ten thousand digits (it is thus used, for example, when looking for large primes in the GIMPS project).

the integers  $\mp 4 \cdot 23 \cdot 14 \pm 3 \cdot 31 \cdot 18 \pmod{713}$ . We thus know that one of the integers

$$386, 603, 110, 327$$

is the message that was sent.  $\square$

**11.G.7.** Show how to encrypt and decrypt the message  $M = 321$  in Rabin cryptosystem with  $n = 437$ .

**Solution.** The encrypted text can be obtained as the square modulo  $n$ :  $C = 321^2 \equiv (-116)^2 = 13456 \equiv 346 \pmod{437}$ . On the other hand, when decrypting, we will use the factorization (its knowledge is the private key of the message receiver)  $n = 437 = 19 \cdot 23$ , and we compute  $r = 346^{\frac{19+1}{4}} = 346^5 \equiv 17 \equiv -2 \pmod{19}$  and  $s = 246^{\frac{23+1}{4}} = 346^6 \equiv 1 \pmod{23}$ . Applying Euclidean algorithm to the pair  $(19, 23) = 1$ , we determine the coefficients into Bézout's identity

$$19 \cdot (-6) + 23 \cdot 5 = 1.$$

Then, the message is one of the integers  $\pm 6 \cdot 19 \cdot 1 \pm 5 \cdot 23 \cdot (-2) \pmod{437}$ , i.e.,  $M = \pm 116$  or  $M = \pm 344$ . Indeed,  $M = -116 \equiv 321 \pmod{437}$ .  $\square$

### 11.6.2. Greatest common divisor and modular inverses.

As we have already shown, the computation of the solution of the congruence  $a \cdot x \equiv 1 \pmod{m}$  in variable  $x$  can be easily reduced (thanks to Bézout's identity) to the computation of the greatest common divisor of the integers  $a$  and  $m$  and looking for the coefficients  $k, l$  in Bézout's identity  $k \cdot a + l \cdot m = 1$  (the integer  $k$  is then the wanted inverse of  $a$  modulo  $m$ ).

```
function extended_gcd(a,m)
  if m == 0:
    return (1,0)
  else
    (q,r) := divide(a,m)
    (k,l) := extended_gcd(m,r)
    return (1,k - q*l)
```

A thorough analysis<sup>8</sup> shows that the problem of computing the greatest common divisor has *quadratic* time complexity.

**11.6.3. Modular exponentiation.** The algorithm for modular exponentiation is based on the idea that when computing, for instance  $2^{64} \pmod{1000}$ , one need not calculate  $2^{64}$  and then divide it with remainder by 1000, but that it is better to multiply the 2's gradually and reduce the temporary result modulo 1000 whenever it exceeds this value. More importantly, there is no need to perform such a huge number of multiplications: in this case, 63 naive multiplications can be replaced with six squarings, as

$$2^{64} = (((((2^2)^2)^2)^2)^2)^2.$$

```
function modular_pow (base, exp, mod)
  result := 1
  while exp > 0
    if (exp % 2 == 1):
      result := (result * base) % mod
    exp := exp >> 1
    base := (base * base) % mod
  return result
```

The algorithm squares the base modulo  $n$  for every binary digit of the exponent (which can be done in quadratic time in the worst case) and it performs a multiplication for every one in the binary representation of the exponent. Altogether, we are able to do modular exponentiation in *cubic* time in the worst case. We can also notice that the complexity is a good deal dependent on the binary appearance of the exponent.

**Example.** Let us compute  $2^{560} \pmod{561}$ .

Since  $560 = (1000110000)_2$ , the mentioned algorithm gives

<sup>8</sup>See, for example, D. Knuth, *Art of Computer Programming, Volume 2: Seminumerical Algorithms*, Addison-Wesley 1997 or Wikipedia, *Euclidean algorithm*, [http://en.wikipedia.org/wiki/Euclidean\\_algorithm](http://en.wikipedia.org/wiki/Euclidean_algorithm) (as of July 29, 2017).

exp	base	result	last digit exp
560	2	1	0
280	4	1	0
140	16	1	0
70	256	1	0
35	460	1	1
17	103	460	1
8	511	256	0
4	256	256	0
2	460	256	0
1	103	256	1
0	511	1	0

Therefore,  $2^{560} \equiv 1 \pmod{561}$ .

**11.6.4. Primality testing.** Although we have the Fundamental theorem of arithmetic, which guarantees that every natural number can be uniquely factored to a product of primes, this operation is very hard from the computational point of view. In practice, it is usually done in the following steps:



- (1) finding all divisors below a given threshold (by trying all primes up to the threshold, which is usually somewhere around  $10^6$ );
- (2) testing the remaining factor for compositeness (deciding whether some necessary condition for primality holds);
  - (a) if the compositeness test did not find the integer to be composite, i.e., it is likely to be a prime, then we test it for primality to verify that it is indeed a prime;
  - (b) if the compositeness test proved that the integer was composite, then we try to find a non-trivial divisor.

The mentioned steps are executed in this order because the corresponding algorithms are gradually (and strongly) increasing in time complexity. In 2002, Agrawal, Kayal, and Saxena published an algorithm for primality testing in polynomial time, but it is still more efficient to use the above procedure in practice.

**11.6.5. Compositeness tests – how to recognize composite numbers with certainty?** The so-called compositeness tests check for some necessary condition for primality. The easiest of such conditions is Fermat’s little theorem.

**Proposition** (Fermat’s test). *Let  $N$  be a natural number. If there is an  $a \not\equiv 0 \pmod{N}$  such that  $a^{N-1} \not\equiv 1 \pmod{N}$ , then  $N$  is not a prime.*

Unfortunately, having a composite  $N$ , it still may not be easy to find such an integer  $a$  which reveals the compositeness of  $N$ . There are even such exceptional integers  $N$  for which the only integers  $a$  with the mentioned property are those which are not coprime to  $N$ . To find them is thus equivalent to finding a divisor, and thus to factoring  $N$  to primes.

There are indeed such ugly (or extremely nice?) composite numbers  $N$  for which every integer  $a$  which is coprime to  $N$  satisfies  $a^{N-1} \equiv 1 \pmod{N}$ . These are called

*Carmichael numbers*, the least of which<sup>9</sup> is  $561 = 3 \cdot 11 \cdot 17$ , and it was no sooner than in 1992 that it was proved<sup>10</sup> that there are even infinitely many of them.

**Example.** We will prove that 561 is a Carmichael number, i.e., that it holds for every  $a \in \mathbb{N}$  which is coprime to  $3 \cdot 11 \cdot 17$  that  $a^{560} \equiv 1 \pmod{561}$ .

Thanks to the properties of congruences, we know that it suffices to prove this congruence modulo 3, 11, and 17. However, this can be obtained straight from Fermat's little theorem since such an integer  $a$  satisfies  $a^2 \equiv 1 \pmod{3}$ ,  $a^{10} \equiv 1 \pmod{11}$ ,  $a^{16} \equiv 1 \pmod{17}$ , where all of 2, 10, and 16 divide 560, hence  $a^{560} \equiv 1$  modulo 3, 11 as well as 17 for all integers  $a$  coprime to 561 (see also Korselt's criterion mentioned below).

**11.6.6. Proposition** (Korselt's criterion). *A composite number  $n$  is a Carmichael number if and only if both of the following conditions hold*

- $n$  is square-free (divisible by the square of no prime),
- $p - 1 \mid n - 1$  holds for all primes  $p$  which divide  $n$ .

PROOF. “ $\Leftarrow$ ” We will show that if  $n$  satisfies the above two conditions and it is composite, then every  $a \in \mathbb{Z}$  which is coprime to  $n$  satisfies  $a^{n-1} \equiv 1 \pmod{n}$ . Let us thus factor  $n$  to the product of distinct odd primes:  $n = p_1 \cdots p_k$ , where  $p_i - 1 \mid n - 1$  for all  $i \in \{1, \dots, k\}$ . Since  $(a, p_i) = 1$ , we get from Fermat's little theorem that  $a^{p_i-1} \equiv 1 \pmod{p_i}$ , whence (thanks to the condition  $p_i - 1 \mid n - 1$ ) it also follows that  $a^{n-1} \equiv 1 \pmod{p_i}$ . This is true for all indices  $i$ , hence  $a^{n-1} \equiv 1 \pmod{n}$ , so  $n$  is indeed a Carmichael number.

“ $\Rightarrow$ ” A Carmichael number  $n$  cannot be even since then we would get for  $a = -1$  that  $a^{n-1} \equiv -1 \pmod{n}$ , which would (since  $a^{n-1} \equiv 1 \pmod{n}$ ) mean that  $n$  is equal to 2 (and thus is not composite). Therefore, let  $n$  factor as  $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ , where  $p_i$  are distinct odd primes and  $\alpha_i \in \mathbb{N}$ . Thanks to theorem 11.3.16, we can choose for every  $i$  a primitive root  $g_i$  modulo  $p_i^{\alpha_i}$ , and the Chinese remainder theorem then yields an integer  $a$  which satisfies  $a \equiv g_i \pmod{p_i^{\alpha_i}}$  for all  $i$  and which is apparently coprime to  $n$ . Further, we know from the assumption that  $a^{n-1} \equiv 1 \pmod{n}$ , so this holds modulo  $p_i^{\alpha_i}$ , and thus  $g_i^{n-1} \equiv 1 \pmod{p_i^{\alpha_i}}$  as well. Since  $g_i$  is a primitive root modulo  $p_i^{\alpha_i}$ , the integer  $n-1$  must be a multiple of its order, i.e. a multiple of  $\varphi(p_i^{\alpha_i}) = p_i^{\alpha_i-1}(p_i - 1)$ . At the same time, we have  $(p_i, n-1) = 1$  (since  $p_i \mid n$ ), so necessarily  $\alpha_i = 1$  and  $p_i - 1 \mid n - 1$ .  $\square$

Fermat's primality test can be slightly improved to Euler's test or even more with the help of the Jacobi symbol, yet this still does not mend the presented problem completely.

<sup>9</sup>The first discoverer of the first seven Carmichael numbers is apparently Czech priest and mathematician Václav Šimerka (1819–1887), who occupied himself with them much earlier than the American mathematician R. D. Carmichael (1879–1967), whose name they bear.

<sup>10</sup>W. R. Alford, A. Granville and C. Pomerance, *There are Infinitely Many Carmichael Numbers*, Annals of Mathematics, Vol. 139, No. 3 (1994), pp. 703–722.

**Proposition** (Euler’s test). *Let  $N$  be an odd natural number. If there is an integer  $a \not\equiv 0 \pmod{N}$  such that  $a^{\frac{N-1}{2}} \not\equiv \pm 1 \pmod{N}$ , then  $N$  is not a prime.*

**PROOF.** This follows directly from Fermat’s theorem and the fact that for  $N$  odd, we have  $a^{N-1} = (a^{\frac{N-1}{2}} - 1)(a^{\frac{N-1}{2}} + 1)$ .  $\square$

**Proposition** (Euler-Jacobi test). *Let  $N$  be an odd natural number. If there is an integer  $a \not\equiv 0 \pmod{N}$  such that  $a^{\frac{N-1}{2}} \not\equiv \left(\frac{a}{N}\right) \pmod{N}$ , then  $N$  is not a prime.*

**PROOF.** This follows immediately from lemma 11.4.12.  $\square$

**Example.** Let us consider  $N = 561 = 3 \cdot 11 \cdot 17$  as before and let  $a = 5$ . Then, we have  $5^{280} \equiv 1 \pmod{3}$  and  $5^{280} \equiv 1 \pmod{10}$ , but  $5^{280} \equiv -1 \pmod{17}$ , so surely  $5^{280} \not\equiv \pm 1 \pmod{561}$ . Here, it did not hold that  $a^{(N-1)/2} \equiv \pm 1 \pmod{N}$ , so we even did not need to check the value of the Jacobi symbol  $(5/561)$ . However, the Euler-Jacobi test can often reveal a composite number even in the case when this power is equal to  $\pm 1$ .

**Example.** Euler’s test cannot detect the compositeness of the integer  $N = 1729 = 7 \cdot 13 \cdot 19$  since the integer  $\frac{N-1}{2} = 864 = 2^5 \cdot 3^3$  is divisible by 6, 12, and 18, and so it follows from Fermat’s theorem that  $a^{(N-1)/2} \equiv 1 \pmod{N}$  holds for all integers  $a$  coprime to  $N$ . On the other hand, we get already for  $a = 11$  that  $(11/1729) = -1$ , so the Euler-Jacobi is able to recognize the integer 1729 as composite.

Let us notice that the value of the Legendre or Jacobi symbol  $(a/n)$  can be computed very efficiently thanks to the law of quadratic reciprocity<sup>11</sup>, namely in time  $O((\log a)(\log n))$ .

#### PSEUDOPRIMES

A composite number  $n$  is called a *pseudoprime* if it passes the corresponding test of compositeness without being revealed. We thus have

- (1) Fermat pseudoprimes to base  $a$ ,
- (2) Euler (or Euler-Jacobi) pseudoprimes to base  $a$ ,
- (3) strong pseudoprimes to base  $a$ , which are composite numbers which pass the following compositeness test:

The subsequent test is simple, yet (as shown in theorem 11.6.8) very efficient. It is a further specification of Fermat’s test, which we have introduced at the beginning.

**11.6.7. Theorem.** *Let  $p$  be an odd prime. Let us write  $p-1 = 2^t \cdot q$ , where  $t$  is a natural number and  $q$  is odd. Then, every integer  $a$  which is not a multiple of  $p$  satisfies  $a^q \equiv 1 \pmod{p}$  or there exists an  $e \in \{0, 1, \dots, t-1\}$  such that  $a^{2^e q} \equiv -1 \pmod{p}$ .*

<sup>11</sup>See H. Cohen, *A Course in Computational Algebraic Number Theory*, Springer, 1993.

PROOF. It follows from Fermat's little theorem that

$$\begin{aligned}
 p \mid a^{p-1} - 1 &= (a^{\frac{p-1}{2}} - 1)(a^{\frac{p-1}{2}} + 1) = \\
 &= (a^{\frac{p-1}{4}} - 1)(a^{\frac{p-1}{4}} + 1)(a^{\frac{p-1}{2}} + 1) = \\
 &\quad \vdots \\
 &= (a^q - 1)(a^q + 1)(a^{2q} + 1) \cdots (a^{2^{t-1}q} + 1),
 \end{aligned}$$

whence the statement follows easily since  $p$  is a prime.  $\square$

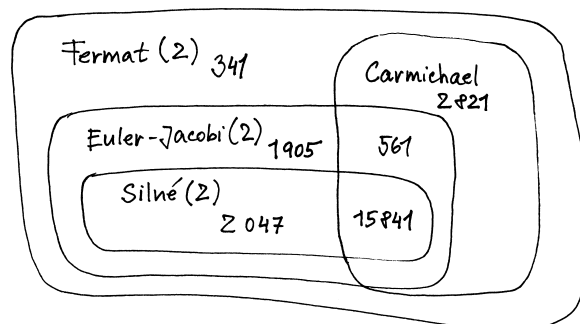
**Proposition** (Miller-Rabin compositeness test). *Let  $N$ ,  $t$ ,  $q$  be natural numbers such that  $N$  is odd and  $N - 1 = 2^t \cdot q$ ,  $2 \nmid q$ . If there is an integer  $a \not\equiv 0 \pmod{N}$  such that*

$$a^q \not\equiv 1 \pmod{N}$$

$$a^{2^e q} \not\equiv -1 \pmod{N} \quad \text{for } e \in \{0, 1, \dots, t-1\},$$

then  $N$  is not a prime.

PROOF. The correctness of the test follows directly from the previous theorem.  $\square$



*Miscellaneous types of pseudoprimes*

In practice, this easy test rapidly increases the ability to recognize composite numbers. The least strong pseudoprime to base 2 is 2047 (while the least Fermat pseudoprime to base 2 was already 341), and considering the bases 2, 3, and 5, the least strong pseudoprime is 25326001. In other words, if we are to test integers below  $2 \cdot 10^7$ , then it is sufficient to execute this compositeness test already for the bases 2, 3, and 5. If the tested integer is not revealed to be composite, then it is surely a prime. On the other hand, it has been proved that no finite basis is sufficient for testing all natural numbers.

The Miller-Rabin test is a practical application of the previous statement, and we are even able to bound the probability of failure thanks to the following theorem, which we present without a proof <sup>12</sup>.

**11.6.8. Theorem.** *Let  $N > 10$  be an odd composite number. Let us write  $N - 1 = 2^t \cdot q$ , where  $t$  is a natural number and  $q$  is odd. Then, at most a quarter of the integers from the set  $\{a \in \mathbb{Z}; 1 \leq a < N, (a, N) = 1\}$  satisfies the following condition:*

$$a^q \equiv 1 \pmod{N}$$

<sup>12</sup>Schoof, René (2004), "Four primality testing algorithms" (PDF), Algorithmic Number Theory: Lattices, Number Fields, Curves and Cryptography, Cambridge University Press, ISBN 978-0-521-80854-5

or there is an  $e \in \{0, 1, \dots, t - 1\}$  satisfying

$$a^{2^e q} \equiv -1 \pmod{N}.$$

In practical implementations, one usually tests about 20 random bases (or the least prime bases). In this case, the above theorem states that the probability of failing to reveal a composite number is less than  $2^{-40}$ .

The time complexity of the algorithm is same as in the case of modular exponentiation, i.e. *cubic* in the worst case. However, we should realize that the test is non-deterministic and the reliability of its deterministic version depends on the so-called generalized Riemann hypothesis (GRH<sup>13</sup>).

**11.6.9. Primality tests.** Primality tests are usually applied



when the used compositeness test claims that the examined integer is *likely* to be a prime, or they are executed straightaway for special types of integers. Let us first give a list of the most known tests, which includes historical tests as well as very modern ones.

- (1) AKS – a general polynomial primality test discovered by Indian mathematicians Agrawal, Kayal, and Saxena in 2002.
- (2) Pocklington-Lehmer test – primality test of subexponential complexity.
- (3) Lucas-Lehmer test – primality test for Mersenne numbers.
- (4) Pépin's test – primality test for Fermat numbers from 1877.
- (5) ECPP - primality test based on the so-called elliptic curves.

Now, we will introduce a standard primality test for Mersenne numbers.

**Proposition** (Lucas-Lehmer test). *Let  $q \neq 2$  be a prime, and let a sequence  $(s_n)_{n=0}^\infty$  be defined recursively by*

$$s_0 = 4, s_{n+1} = s_n^2 - 2.$$

*Then, the integer  $M_q = 2^q - 1$  is a prime if and only if  $M_q$  divides  $s_{q-2}$ .*

**PROOF.** We will be working in the ring  $R = \mathbb{Z}[\sqrt{3}] = \{a + b\sqrt{3}; a, b \in \mathbb{Z}\}$ , where the division with remainder behaves similarly as in the integers (see also 12.3.5). Let us set  $\alpha = 2 + \sqrt{3}$ ,  $\beta = 2 - \sqrt{3}$  and note that  $\alpha + \beta = 4$ ,  $\alpha \cdot \beta = 1$ .

First, we prove by induction that it holds for all  $n \in \mathbb{N}_0$  that

$$(1) \quad s_n = \alpha^{2^n} + \beta^{2^n} = \beta^{2^n} \left( 1 + \alpha^{2^{n+1}} \right).$$

The statement is true for  $n = 0$  since  $s_0 = 4 = \alpha + \beta$ . Now, let us suppose that it is true for  $n - 1$ , then  $s_n = s_{n-1}^2 - 2$  is, by the induction hypothesis, equal to  $\left( \alpha^{2^{n-1}} + \beta^{2^{n-1}} \right)^2 - 2 = \alpha^{2^n} + \beta^{2^n}$ .

<sup>13</sup>Wikipedia, *Riemann hypothesis*, [http://en.wikipedia.org/wiki/Riemann\\_hypothesis](http://en.wikipedia.org/wiki/Riemann_hypothesis) (as of July 29, 2017).

Further, since  $M_q \equiv -1 \pmod{8}$ , we have  $(2/M_q) = 1$ , and it follows from the law of quadratic reciprocity that

$$\left(\frac{3}{M_q}\right) = -\left(\frac{M_q}{3}\right) = -\left(\frac{2^q - 1}{3}\right) = -\left(\frac{1}{3}\right) = -1,$$

since we have  $2^q - 1 \equiv 1 \pmod{3}$  for  $q$  odd. Both of these expressions are valid even if  $M_q$  is not a prime (in this case, it is the Jacobi symbol).

Let us note that in the last part of the proof, we will use the extension of the congruence relation to the elements of the domain  $\mathbb{Z}[\sqrt{3}] = \{a+b\sqrt{3}; a, b \in \mathbb{Z}\}$ ; just like in the case of the integers, we write for  $\alpha, \beta \in \mathbb{Z}[\sqrt{3}]$  that  $\alpha \equiv \beta \pmod{p}$  if  $p \mid \alpha - \beta$ . Further, an analog of proposition (ii) from 11.C.6 holds as well – if  $p$  is a prime, then  $(\alpha + \beta)^p \equiv \alpha^p + \beta^p \pmod{p}$  (the proof is identical to the one for the integers).

“ $\implies$ ” Suppose that  $M_q$  is a prime. We will prove that  $\alpha^{2^{q-1}} \equiv -1 \pmod{M_q}$ , which will imply (thanks to 1) that  $M_q \mid s_{q-2}$ . Since  $2^{(M_q-1)/2} \equiv (2/M_q) = 1 \pmod{M_q}$ , there is a  $y \in \mathbb{Z}$  such that  $2y^2 \equiv 1 \pmod{M_q}$ . We have

$$(y(1 + \sqrt{3}))^2 = y^2(4 + 2\sqrt{3}) \equiv \alpha \pmod{M_q},$$

whence, invoking Fermat's theorem and the relation  $2^{q-1} = \frac{M_q+1}{2}$ , we get

$$\begin{aligned} \alpha^{2^{q-1}} &\equiv \left(y(1 + \sqrt{3})\right)^{M_q+1} \\ &\equiv y^2 \cdot y^{M_q-1} (1 + \sqrt{3}) \cdot (1 + \sqrt{3})^{M_q} \\ &\equiv y^2 (1 + \sqrt{3}) \cdot (1 - \sqrt{3}) = -2y^2 \\ &\equiv -1 \pmod{M_q}. \end{aligned}$$

When deriving this, we made use of the fact that 3 is a quadratic nonresidue modulo  $M_q$ , so

$$\begin{aligned} (1 + \sqrt{3})^{M_q} &\equiv 1 + (\sqrt{3})^{M_q} = 1 + 3^{(M_q-1)/2} \cdot \sqrt{3} \\ &\equiv 1 - \sqrt{3} \pmod{M_q}. \end{aligned}$$

“ $\impliedby$ ” For the other direction, let  $M_q \mid s_{q-2}$ . However, then

$$M_q \mid s_{q-2} \cdot \alpha^{2^{q-2}} = 1 + \alpha^{2^{q-1}}.$$

If  $p \neq 2, 3$  is a prime divisor of  $M_q$ , then  $\alpha^{2^{q-1}} \equiv -1 \pmod{p}$  as well and  $\alpha^{2^q} \equiv 1 \pmod{p}$ . Hence it follows that  $2^q$  is the order of  $\alpha$  in the multiplicative group  $T_p = \{a + b\sqrt{3}; 0 \leq a, b < p\} \setminus \{0\}$ .

If we had  $(3/p) = 1$ , then we would get

$$\begin{aligned} \alpha^{p-1} &= \beta \cdot \alpha^p \equiv \beta \cdot \left(2^p + (\sqrt{3})^p\right) \\ &\equiv \beta \cdot \left(2 + \sqrt{3} \cdot 3^{(p-1)/2}\right) \equiv \beta \cdot (2 + \sqrt{3}) = 1, \end{aligned}$$

whence  $p - 1$  would be a multiple of the order of  $\alpha$ , i.e.  $2^q$ . However, this would mean that  $p > p - 1 \geq 2^q > 2^q - 1 = M_q$ , which contradicts the fact that  $p$  is a divisor of  $M_q$ .



Therefore, we have  $(3/p) = -1$  and

$$\begin{aligned}\alpha^{p+1} &\equiv (2 + \sqrt{3}) (2 + \sqrt{3})^p \\ &\equiv (2 + \sqrt{3}) (2 - \sqrt{3}) \\ &\equiv 1 \pmod{p}.\end{aligned}$$

The order of  $\alpha$  modulo  $p$  is  $2^q$ , hence  $2^q \mid p+1$  and especially  $p \geq 2^q - 1 = M_q$ . At the same time,  $p$  is a prime divisor of  $M_q$ , therefore,  $M_q = p$  is a prime.  $\square$

Unlike the proof, implementation of this algorithm is very easy.



**Algorithm (Lucas-Lehmer primality test):**

```
function LL_is_prime(q)
  s := 4; M := 2q - 1
  repeat q - 2 times
    s := s2 - 2 (mod M)
  if s = 0 return PRIME.
  else return COMPOSITE.
```

The time complexity of this test is asymptotically the same as in the case of the Miller-Rabin test. It is, however, more efficient in concrete instances.

*Fermat numbers* are integers of the form  $F_n = 2^{2^n} + 1$ .



Pierre de Fermat conjectured in the 17th century that all of the integers of this form are primes (apparently driven by the effort to generalize the observation for  $F_0 = 3$ ,  $F_1 = 5$ ,  $F_2 = 17$ ,  $F_3 = 257$ , and  $F_4 = 65537$ ). However, in the 18th century, Leonhard Euler found out that  $F_5 = 641 \times 6700417$ , and we have not been able to discover any other Fermat primes so far. Since their size increases rapidly, it takes much time resources to compute with them (so the following test is not much used). Nowadays, the least Fermat number which has not been tested is  $F_{33}$ , which has 2 585 827 973 digits, and it is thus much greater than the largest discovered prime.

**Proposition** (Pépin's test). *A necessary and sufficient condition for the  $n$ -th Fermat number  $F_n$  to be a prime is*

$$3^{\frac{F_n - 1}{2}} \equiv -1 \pmod{F_n}.$$

We can see that this is a very simple test, which is actually a mere part of Euler's compositeness test.

**PROOF OF CORRECTNESS OF PÉPIN'S TEST.** First, suppose that  $3^{(F_n - 1)/2} \equiv -1 \pmod{F_n}$ . Then,  $3^{F_n - 1} \equiv 1 \pmod{F_n}$ . Since  $F_n - 1$  is a power of two,  $F_n - 1$  is necessarily the order of 3 modulo  $F_n$ . However, the order of every integer modulo  $F_n$  is at most  $\varphi(F_n) \leq F_n - 1$ , hence in this case, we have  $\varphi(F_n) = F_n - 1$ , which means that  $F_n$  is a prime.

For the other direction, let  $F_n$  be a prime. From part (i) of lemma 11.4.12, we get that  $3^{(F_n-1)/2} \equiv (3/F_n) \pmod{F_n}$ , so it suffices to determine the value  $(3/F_n)$ . However, this is easy, because  $F_n \equiv 2 \pmod{3}$ , and thus  $(F_n/3) = -1$ . Further, we have  $F_n \equiv 1 \pmod{4}$ , and the law of quadratic reciprocity thus yields  $(3/F_n) = -1$ , which is what we wanted to prove.  $\square$

Now, we will introduce a primality test which is a bit old, yet it is still widely used in modern computation systems – the so-called *Pocklington-Lehmer* test. However, first of all, we will describe a simpler primality test for illustration, the so-called *Lucas's test*:



**11.6.10. Theorem (Lucas).** *If for all prime divisors  $q$  of  $N - 1$ , there is an  $a$  such that  $a^{N-1} \equiv 1 \pmod{N}$ ,  $a^{\frac{N-1}{q}} \not\equiv 1 \pmod{N}$ , then  $N$  is a prime.*

**PROOF.** It suffices to prove that  $N - 1$  divides  $\varphi(N)$  (which is a condition apparently unsatisfied by composite numbers). If not, then there is a prime  $q$  and  $r \in \mathbb{N}$  such that  $q^r$  divides  $N - 1$ , but it does not divide  $\varphi(N)$ . The order  $e$  of the integer  $a$  divides  $N - 1$  (the first condition) and does not divide  $(N - 1)/q$  (the second condition), hence  $q^r$  divides  $e$ . Furthermore,  $e$  divides  $\varphi(N)$ , and so  $q^r$  does, a contradiction.  $\square$

The integer  $a$  from the previous theorem is called a *primality witness* for the integer  $N$  (in this as well as in the other primality tests).

Another general primality test is based on the above one. It is good if we want to make the high probability of the answer of the Miller-Rabin compositeness test into certainty.

**11.6.11. Theorem (Pocklington–Lehmer).** *Let  $N$  be a natural number,  $N > 1$ . Let  $p$  be a prime which divides  $N - 1$ . Further, let us suppose that there is an integer  $a_p$  such that*

$$a_p^{N-1} \equiv 1 \pmod{N} \quad \text{and} \quad \left( a_p^{\frac{N-1}{p}} - 1, N \right) = 1.$$

*Let  $p^{\alpha_p}$  be the highest power of  $p$  which divides  $N - 1$ . Then every positive divisor  $d$  of the integer  $N$  satisfies*

$$d \equiv 1 \pmod{p^{\alpha_p}}.$$

**PROOF OF THE POCKLINGTON–LEHMER THEOREM.** Every positive divisor  $d$  of the integer  $N$  is a product of prime divisors of  $N$ , so it suffices to prove the theorem for prime values of  $d$ . The condition  $a_p^{N-1} \equiv 1 \pmod{N}$  implies that the integers  $a_p, N$  are coprime (any divisor they have in common must divide the right-hand side of the congruence as well). Then,  $(a_p, d) = 1$  as well, and we have  $a_p^{d-1} \equiv 1 \pmod{d}$  by Fermat's theorem. Since  $(a_p^{(N-1)/p} - 1, N) = 1$ , we get  $a_p^{(N-1)/p} \not\equiv 1 \pmod{d}$ .

Let  $e$  denote the order of  $a_p$  modulo  $d$ . Then,  $e \mid d - 1$ ,  $e \mid N - 1$ , and  $e \nmid (N - 1)/p$ .

If  $p^{\alpha_p} \nmid e$ , then  $e \mid N - 1$  would imply that  $e \mid \frac{N-1}{p}$ , which is a contradiction. Therefore,  $p^{\alpha_p} \mid e$ , and so  $p^{\alpha_p} \mid d - 1$ .  $\square$



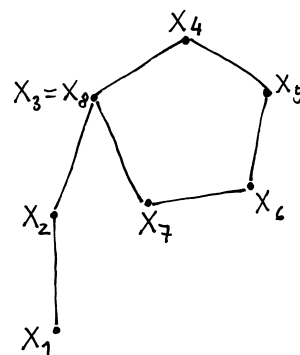


**11.6.14. Looking for divisors.** If one of the compositeness tests verifies that a given integer is indeed composite, we usually want to find one of its non-trivial divisors. However, this task is much more difficult than mere revealing that it is composite – let us recall that the compositeness tests can guarantee the compositeness, yet they provide us with no divisors (which is, on the other hand, advantageous for RSA and similar cryptographic protocols). Therefore, we will present here only a short summary of methods used in practice and one sample for inspiration.



- (1) Trial division
- (2) Pollard's  $\rho$ -algorithm
- (3) Pollard's  $p - 1$  algorithm
- (4) Elliptic curve method (ECM)
- (5) Quadratic sieve (QS)
- (6) Number field sieve (NFS)

For illustration, we demonstrate in the exercises (11.F.8) one of these algorithms – Pollard's  $\rho$ -method – on a concrete instance. This algorithm is especially suitable for finding *relatively* small divisors (since its expected complexity depends on the size of these divisors), and it is based on the idea that having a random function  $f : S \rightarrow S$ , where  $S$  is a finite set having  $n$ -elements, the sequence  $(x_n)_{n=0}^{\infty}$ , where  $x_{n+1} = f(x_n)$ , must loop. The expected length of the tail as well as the period is then  $\sqrt{\pi \cdot n/8}$ .



The algorithm described below is again a straightforward implementation of the mentioned reasonings.

**Algorithm (Pollard's  $\rho$ -method):**

Input:  $n$  — the integer to be factored,  
and an appropriate function  $f(x)$

$a := 2; b := 2; d := 1$

While  $d = 1$  do

$a := f(a)$

$b := f(f(b))$

$d := \gcd(a - b, n)$

If  $d = n$ , return FAILURE.

Else return  $d$ .

**11.6.15. Public-key cryptography.** In present-day practice, the most important application of number theory is the so-called public-key cryptography. We shall only briefly touch this fascinating area of Mathematics.



The main objectives are to provide

- encryption; the message *encrypted* with the public key of the receiver can be decrypted by no one else (to be precise, by no one who does not know his private key);
- signature; the integrity of the message *signed* with the private key of the sender can be verified by anyone with access to his public key.

The most basic and most often used protocols in public-key cryptography are:

- *RSA* (encryption) and the derived system for signing messages,
- *Digital Signature Algorithm* – DSA and its variant based on elliptic curves (ECDSA),
- *Rabin cryptosystem* (and signature scheme),
- *ElGamal cryptosystem* (and signature scheme),
- *elliptic curve cryptography* (ECC),
- *Diffie-Hellman key exchange protocol* (DH).

**11.6.16. Encryption – RSA.** First, we describe the most known public-key cipher – RSA. The simple principle of the protocol *RSA*<sup>14</sup> is based on the Euler’s theorem 11.3.12:

- The user *A* needs a pair of keys – a public one ( $V_A$ ) and a private one ( $S_A$ ).
- Key generating: the user selects two large primes  $p, q$ , and calculates  $n = pq$ ,  $\varphi(n) = (p - 1)(q - 1)$ . The integer  $n$  is part of the public key; the main idea is that it is too hard to compute  $\varphi(n)$  without knowing the decomposition.
- Then, the user chooses the second half of the *public key*,  $e$ , and verifies that  $(e, \varphi(n)) = 1$ .
- Using the Euclidean algorithm, the *private key*  $d$  is computed so that  $e \cdot d \equiv 1 \pmod{\varphi(n)}$ .

THE PRINCIPLE OF RSA

The secret communication runs in the following steps (for the sake of simplicity, we will further identify the encryption procedure with the public key  $V_A$  and the decryption procedure with the private key  $S_A$ ):

- Encrypting a numerical *code* of a message  $M$  addressed to the user  $A$  (by any other participant who has got access to the public key  $V_A$ ):

$$C = V_A(M) \equiv M^e \pmod{n}.$$

- Decrypting the cipher  $C$  by the user  $A$ :

$$OT = S_A(C) \equiv C^d \pmod{n}.$$

<sup>14</sup>Ron Rivest, Adi Shamir, Leonard Adleman (1977); C. Cocks, the secret service GCHQ (not publicly) as early as 1973.

The proof of correctness of this protocol (i.e., that  $A$  indeed receives what was meant) is a straightforward application of Euler's theorem: Thanks to 11.3.13, it holds for any message  $M$  which is coprime to  $n$  that  $(M^e)^d \equiv M^1 = M \pmod{n}$ . In the (extremely unlikely) case that the message  $M$  would not be coprime to  $n$ , the statement holds as well, although the proof needs to be modified with the help of the Chinese remainder theorem (however, we should realize that if the message  $M$  with property  $0 < M < n$  is not coprime to  $n$ , then it means that  $(M, n)$  is a non-trivial divisor of  $n$ , so the key of the receiver is actually discredited).

The security of RSA has been tested since it was invented in 1977, and no meaningful weakness (except for side channels or some singular keys) has been discovered yet (provided a sufficiently large key is used; nowadays it is recommended to use at least 2048 bits). Nevertheless, it has not been proved that the RSA problem *really* relies on the hardness of integer factorization.

The requirements on a secure choice of the key for practical reasons are:

- $d$  is large enough (defense against the so-called Wiener's attack),
- $p$  and  $q$  are not too close to each other (see the exercise 11.G.1),
- the public key is selected to be at least  $e = 65537$  (although no direct attack against a small public key  $e$  has been noticed).

**11.6.17. Rabin cryptosystem.** Further, we mention a simplified variant of the protocol named *Rabin cryptosystem*<sup>15</sup>, which has been the first public cryptosystem where one demonstrably needs to factorize the modulus to break it (unlike RSA, for which this has not been proved):



Every participant  $A$  needs a pair of keys – a public one ( $V_A$ ) and a private one ( $S_A$ ).

- Key generating:  $A$  chooses two large primes of roughly the same size –  $p, q \equiv 3 \pmod{4}$ , and computes  $n = pq$ .
- The public key is  $V_A = n$ , the private key is the pair  $S_A = (p, q)$ .

The secret communication then runs as follows:

- Encryption of the numerical *code* of the message  $M$ :  $C = V_A(M) \equiv M^2 \pmod{n}$ .
- Decryption of the cipher  $C$ : the (four) roots of  $C$  modulo  $n$  are computed and it is easily found out which one of them is the original message (for instance, the other three make no sense or do not contain the agreed identification).

As we can see from the description of the protocol, the process of decryption requires the *computation of the square*

<sup>15</sup>Rabin, Michael. *Digitalized Signatures and Public-Key Functions as Intractable as Factorization* (in PDF). MIT Laboratory for Computer Science, January 1979.

root of  $C$  modulo  $n = pq$ , where  $p \equiv q \equiv 3 \pmod{4}$ . This can be done as follows:

- The values  $r \equiv C^{(p+1)/4} \pmod{p}$  and  $s \equiv C^{(q+1)/4} \pmod{q}$  are computed.
- Further, we need to determine the coefficients  $a, b$  in Bézout's identity, i.e., integers for which  $ap + bq = 1$ .
- We set  $x \equiv (aps + bqr) \pmod{n}$ ,  $y \equiv (aps - bqr) \pmod{n}$ .
- The square roots of  $C$  modulo  $n$  are then  $\pm x, \pm y$ .

Let us mention that this is in fact an application of the Chinese remainder theorem and the fact that we are able to easily find the solutions of the quadratic congruence  $x^2 \equiv a \pmod{p}$  provided  $p \equiv 3 \pmod{4}$  (see exercise 11.D.38). Indeed, it holds that

$$\begin{aligned} (\pm x)^2 &= (aps + bqr)^2 \equiv (bqr)^2 \\ &\equiv r^2 \equiv C^{(p+1)/2} \equiv C \pmod{p}, \end{aligned}$$

where we made use of the fact that  $bq \equiv 1 \pmod{p}$  and that  $C \equiv M^2 \pmod{p}$  is a quadratic residue modulo  $p$ , hence  $C^{(p-1)/2} \equiv (C/p) = 1 \pmod{p}$ . Similarly, we have  $(\pm x)^2 \equiv C \pmod{q}$  as well, thus  $\pm x$  is a square root of  $C$  modulo  $n$ . The derivation of the same result for  $y$  is nearly identical.

**11.6.18. Digital signature.** Now, let us briefly describe the principle of digital signature.

PRINCIPLE OF THE DIGITAL SIGNATURE

**Creating the signature:**

- (1) A digest (hash)  $H_M$  of the message is generated, the length of the hash is fixed (160 or 256 bits, for instance) – we should realize that such a mapping is surely not injective (there will be many messages sharing the same hash).
- (2) The signature of the message  $S_A(H_M)$  is created from this hash using the knowledge of the private key of the signer (similarly to decryption of a message's text).
- (3) The message  $M$  is sent (optionally encrypted with the public key of the receiver) together with the created signature.

The **signature verification** then runs as follows:

- (1) A digest  $H'_M$  is generated for the received message  $M$  (after decryption, if it has been encrypted)
- (2) Using the public key of the (declared) sender of the message, the original digest of the message is reconstructed:  $V_A(S_A(H_M)) = H_M$ .
- (3) The digests are then compared; i.e., it is found out whether  $H_M = H'_M$ .

The (cryptographic) hash function mentioned above must have the following properties:

- It is easy to find the hash of any message.



- It is impossible (in real time) to find (any) message with the desired hash.
- It is impossible (in real time) to find two messages with the same hash (the function must be *collision-resistant*).
- Every change of the message changes the hash as well.

The most known examples of such functions are:

- MD5 (128 bit, Rivest 1992) – not collision-resistant
- SHA-1 (160 bit, NSA 1995) – from 2005 considered insufficiently collision-resistant
- RIPEMD-320
- SHA-3

**11.6.19. Diffie-Hellman key exchange system.** Another important type of protocol, which is very often used in practice, is a *protocol for key exchange in symmetric cryptography – Diffie-Hellman key exchange*<sup>16</sup>, whose discovery was a breakthrough in this discipline, making it possible to replace one-time keys, messengers with cases (and similar) with mathematical means, in particular without the necessity of prior communication of both sides.

The protocol for the agreement of two sides (Alice, Bob) on a common key (integer) is as follows:

PRINCIPLE OF DH KEY EXCHANGE PROTOCOL

- Both sides agree on a prime  $p$  and a primitive root  $g$  modulo  $p$  (this need not be done secretly).
- Alice chooses a random integer  $a$  and sends  $g^a \pmod{p}$ .
- Bob chooses a random integer  $b$  and sends  $g^b \pmod{p}$ .
- The common key for the communication is then  $g^{ab} \pmod{p}$ .

The security of this protocol relies on the fact that it is hard to compute the discrete logarithm (the so-called *discrete logarithm problem*) – see also part 11.3.15.

There is another encryption algorithm which is based on the Diffie-Hellman key exchange protocol – *algorithm ElGamal*, which we will also describe in short:

- Every participant chooses a prime  $p$  with a primitive root  $g$ .
- Further, they choose a *private key*  $x$ , compute  $h = g^x \pmod{p}$ , and publish the *public key*  $(p, g, h)$ .

The secret communication then runs as follows:

- Encryption of the numerical code of the message  $M$ : choosing a random  $y$  and computing  $C_1 = g^y \pmod{p}$  and  $C_2 = M \cdot h^y \pmod{p}$ , then sending the pair  $(C_1, C_2)$  to participant  $A$ .
- The participant  $A$  then decrypts the message by computing  $C_2/C_1^x$ .

<sup>16</sup> Whitfield Diffie, Martin Hellman (1976); M. Williamson (secret service GCHQ) as early as 1974 (not published).

**Remark.** The mechanism of digital signature can be derived from the ElGamal algorithm just like in the case of RSA.

**H. Additional exercises to the whole chapter**

- 11.H.1.** Find all three-digit  $n \in \mathbb{N}$ , for which  $11 \mid n$  and  $\frac{n}{11}$  is the sum of squares of the digits of  $n$ .
- 11.H.2.** Someone falsely memorized Fermat's little theorem, and claims that for  $p \nmid a$  we have  $a^p \equiv 1 \pmod{p}$ . Find all pairs  $(a, p)$  for which this is true.
- 11.H.3.** Prove that last four digits of  $5^n$ , ( $n = 4, 5, 6, \dots$ ) form a periodical sequence. Find this period.
- 11.H.4.** Prove that there are infinitely many odd natural numbers  $k$  such that the integer  $2^{2^n} + k$  is composite for every  $n \in \mathbb{N}$ .
- 11.H.5.** Prove that for every integer  $k \neq 1$ , there are infinitely many natural numbers  $n$  such that the integer  $2^{2^n} + k$  is composite.
- 11.H.6.** Consider the sequence  $(a_n)_{n=1}^{\infty}$  defined by
- $$a_n = 2^n + 3^n + 6^n - 1.$$
- Prove that for every prime  $p$ , this sequence contains a multiple of  $p$ .
- 11.H.7.** Prove that no natural number  $n$  greater than 1 satisfies  $n \mid 2^n - 1$ .
- 11.H.8.** Prove that for every odd prime  $p$ , there are infinitely many natural numbers  $n$  such that  $p \mid n \cdot 2^n + 1$ .
- 11.H.9.** Let a function  $f : \mathbb{N} \rightarrow \mathbb{N}$  satisfy  $(f(a), f(b)) = (f(a), f(|a - b|))$  for all  $a, b \in \mathbb{N}$ . Prove that  $(f(a), f(b)) = f((a, b))$ . Show that this implies the result of exercise 11.A.7 as well as the fact that  $(F_a, F_b) = F_{(a,b)}$ , where  $F_a$  denotes the  $a$ -th term of the Fibonacci sequence.
- 11.H.10.** Let the RSA parameters be  $n = 143 = 11 \cdot 13$ ,  $e = 7$ ,  $d = 103$ . Sign the message  $m = 8$ , and verify this signature. Decide, whether  $s = 42$  is the signature of the message  $m = 26$ .

## Key to the exercises

**10.C.10.** The answer is  $\frac{3}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot 1 = \frac{4}{5}$ .

**10.E.17.** Simply,  $a = \frac{3}{8}$ . Thus, the distribution function of the random variable  $X$  is  $F_X(t) = \frac{1}{8}t^3$  for  $t \in (0, 2)$ , zero for smaller values of  $t$ , and one for greater. Let  $Z = X^3$  denote the random variable corresponding to the volume of the considered cube. It lies in the interval  $(0, 8)$ . Thus, for  $t \in (0, 8)$  and the distribution function  $F_Z$  of the random variable  $Z$ , we can write  $F_Z(t) = P[Z < t] = P[X^3 < t] = P[X < \sqrt[3]{t}] = F_X(\sqrt[3]{t}) = \frac{1}{8}t$ . Then, the density is  $f_Z(t) = \frac{1}{8}$  on the interval  $(0, 8)$  and zero elsewhere. Since this is the uniform distribution on the given interval, the expected value is equal to 4.

**10.E.21.**  $f_{(X,Y)}(u, v) = uv$ , for  $(u, v) \in (0, 1) \times (0, 2)$ ,  $f_{(X,Y)}(u, v) = 0$  otherwise. Then, the wanted probability is

$$P = \int_0^1 \int_{x^2}^2 xy \, dy \, dx = \frac{11}{12}.$$

**10.E.22.**  $P = \int_0^{\sqrt[3]{2}} \int_{x^3}^2 xy \, dy \, dx = \frac{\sqrt[3]{4}}{12}$ .

**10.F.9.**  $EU = 1 \cdot 0.6 + 2 \cdot 0.4 = 1.4$ ,  $EU^2 = 0.4 + 4 \cdot 0.6 = 2.8$ ,  $EV = 0.4 + 0.6 + 1.2 = 2.1$ ,  $EV^2 = 0.3 + 1.2 + 3.6 = 5.1$ ,  $E(UV) = 2.8$ ,  $\text{var}(U) = 2.8 - 1.4^2 = 2.8 - 1.96 = 0.84$ ,  $\text{var}(V) = 5.1 - 4.41 = 0.69$ ,  $\text{cov}(UV) = 2.8 - 1.4 \cdot 2.1 = -0.14$ ,  $\rho_{U,V} = \frac{-0.14}{\sqrt{0.84 \cdot 0.69}}$ .

**10.F.10.**  $EX = 1/3$ ,  $\text{var}^2 X = 4/45$ .

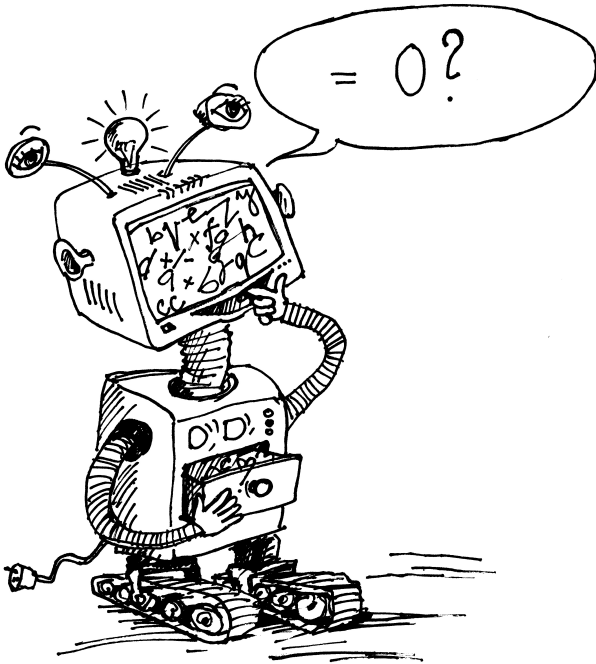
**10.F.11.**

$$\rho_{X,Y} = -1.$$

**10.F.12.**  $\rho_{U,V} \doteq -0,421$ .

## Algebraic structures

*The more abstraction, the more chaos?  
– no, it is often the other way round...*



### A. Boolean algebras and lattices

**12.A.1.** Find the (complete) disjunctive normal form of the proposition

$$(B \Rightarrow C) \wedge [(A \vee C) \wedge B]'$$

#### Solution.

If the propositional formula contains only a few variables (in our case, it is three), the most advantageous procedure is to build the truth table of the formula and build the disjunctive normal form from that. The table will consist of  $2^3 = 8$  rows. The examined formula is denoted  $\varphi$ .



In this chapter, we begin a seemingly very formal study. But the concepts reflect many properties of things and phenomena surrounding us. This is one of the parts of the book which is not in the prerequisites of any other chapter. Large parts serve as a quick illustration of interesting uses of mathematical tools and models.

The simplest properties of real objects are used for encoding in terms of algebraic operations. Thus, “algebra” considers algorithmic manipulations with letters which usually correspond to computations or descriptions of processes.

Strictly speaking, this chapter builds only on the first and sixth parts of chapter one, where abstract views on numbers and relations between objects are introduced. But it is a focal point for abstract versions of many concepts already met.

The first two sections aim at direct generalizations of the familiar algebraic structure of numbers. This leads to a discussion of rings of polynomials. Only then we provide an introduction to group theory, for which there is only a single operation.

The last two sections provide some glimpses of direct applications. The construction of (self-correcting) codes often used in data transfer is considered. The last section explains the elementary foundations of computer algebra. This includes solving polynomial equations and algorithmic methods for manipulation and calculations with formal expressions.

### 1. Posets and Boolean algebras

Familiarity with the properties of addition and multiplication of scalars and matrices is assumed. Likewise, the binary operations of set intersection and union in elementary set theory, as indicated in the end of the first chapter.



We proceed to work with symbols which stand for miscellaneous objects resulting in the universal applicability of the results.

This allows the relating of the basic set operations, to propositional logic which formalizes methods for expressing propositions and evaluating truth values.

**12.1.1. Algebraic operations.** For any set  $M$ , there is a set  $K = 2^M$  consisting of all subsets of  $M$ , together with the operations of union  $\vee : K \times K \rightarrow K$  and intersection  $\wedge :$

A	B	C	$B' \Rightarrow C$	$[(A \vee C) \wedge B]'$	$\varphi$
0	0	0	0	1	0
0	0	1	1	1	1
0	1	0	1	1	1
0	1	1	1	0	0
1	0	0	0	1	0
1	0	1	1	1	1
1	1	0	1	0	0
1	1	1	1	0	0

The resulting complete disjunctive normal form is the disjunction of the formula that correspond to the rows with one in the last column (the formula is true for the given valuation of the atomic propositions). The row corresponds to conjunction of the variables (if the corresponding value is 1) or their negations (if it is 0). In our case, it is the disjunction of conjunctions corresponding to the second, third, and sixth rows, i. e., the result is

$$(A' \wedge B' \wedge C) \vee (A' \wedge B \wedge C') \vee (A \wedge B' \wedge C).$$

We can also rewrite the formula by expanding the connective  $\Rightarrow$  with  $\wedge$  and  $\vee$ , using the De Morgan laws and distributivity:

$$\begin{aligned} & (B' \Rightarrow C) \wedge [(A \vee C) \wedge B] \iff \\ \iff & (B \vee C) \wedge [(A \vee C)' \vee B'] \iff \\ \iff & (B \vee C) \wedge [(A' \wedge C') \vee B'] \iff \\ \iff & [(B \vee C) \wedge (A' \wedge C')] \vee [(B \vee C) \wedge B'] \iff \\ \iff & [(B \wedge A' \wedge C') \vee (C \wedge A' \wedge C')] \vee [(B \wedge B') \vee (C \wedge B')] \iff \\ \iff & (B \wedge A' \wedge C') \vee (C \wedge B'), \end{aligned}$$

which is an (incomplete) disjunctive normal form of the given formula. Clearly, it is equivalent to our result above (the word “complete” means that each disjunct (called *clause* in this context) contains each of the three variables or their negations (these are called *literals*)).  $\square$

**12.A.2.** Find a disjunctive normal form of the formula

$$((A \wedge B) \vee C)' \wedge (A' \vee (B \wedge C' \wedge D))$$

We know several logical connectives:  $\wedge, \vee, \implies, \equiv$  and the unary  $'$ . Any propositional formula with these connectives can be equivalently written using only some of them, for instance  $\vee$  and  $'$ . There are also connectives which alone suffice to express any propositional formula. From binary connectives, these are NAND and NOR ( $A \text{ NAND } B = (A \wedge B)'$ ,  $A \text{ NOR } B = (A \vee B)'$ ). Try to express each of the known connectives using only NAND, and then only NOR. These connectives are implemented in electric circuits as the so-called “gates”.

**12.A.3.** Express the propositional formula  $(A \Rightarrow B)$  using only the NAND gates.  $\circ$

$K \times K \rightarrow K$ . This is an instance of an algebraic structure on the set  $K$  with two *binary operations*. In general, write  $(K, \vee, \wedge)$ . In the special case of sets, these binary operations are denoted rather by  $\cup$  and  $\cap$ , respectively.

To every set  $A \in K$ , its complement  $A' = K \setminus A$  can be assigned. This is another operation  $' : K \rightarrow K$  with only one argument. Such operations are called *unary operations*.

In general, there are algebraic structures with  $k$  operations  $\mu_1, \dots, \mu_k$ , each of them

$$\mu_j : K \times \dots \times K \rightarrow K$$

with  $i_j$  arguments, and write  $(K, \mu_1, \dots, \mu_k)$  for such a structure. The number  $i_j$  of arguments is called the *parity* of the operation (“unary”, “binary”, etc.). If  $i_j = 0$ , then the operation has no arguments which means it is a distinguished element in  $K$ .

With subsets in  $K = 2^M$ , there is the unique “greatest object”, i.e. the entire set  $M$ , which is neutral for the  $\wedge$  operation. Similarly, the empty set  $\emptyset \in K$  is the only neutral element for  $\vee$ . Notice that if  $M$  is empty, then  $K$  contains the only element  $\emptyset$ .

**12.1.2. Set algebra.** View the algebraic structure on the set  $K = 2^M$  from the previous paragraph as  $(K, \vee, \wedge, ', 1, 0)$ , with two binary operations, one unary operation (the complement), and two special elements  $1 = M, 0 = \emptyset$ .

It is easily verified that all elements  $A, B, C \in K$  satisfy the following properties:

AXIOMS OF BOOLEAN ALGEBRAS

- (1)  $A \wedge (B \wedge C) = (A \wedge B) \wedge C,$
- (2)  $A \vee (B \vee C) = (A \vee B) \vee C,$
- (3)  $A \wedge B = B \wedge A, \quad A \vee B = B \vee A,$
- (4)  $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C),$
- (5)  $A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C),$
- (6) there is a  $0 \in K$  such that  $A \vee 0 = A,$
- (7) there is a  $1 \in K$  such that  $A \wedge 1 = A,$
- (8)  $A \wedge A' = 0, \quad A \vee A' = 1.$



Compare these properties with those of the scalars  $(\mathbb{K}, +, \cdot, 0, 1)$ : Properties (1) and (2) say that both the operations  $\wedge$  and  $\vee$  are associative. Property (3) says that both operations are also commutative. So far, this is the same as for the addition and multiplication of scalars. Also there are neutral elements for both operations there.

However, the properties (4) and (5) are stronger now: they require the distributivity of  $\wedge$  over  $\vee$  as well as  $\vee$  over  $\wedge$ . Of course, this cannot be the case for addition and multiplication of numbers. In the case of numbers, multiplication distributes over addition but not vice versa.

**12.A.4.** Write down a logic table for the Boolean proposition

$$((A \wedge B) \vee C)'$$

**Solution.** Using de Morgan's and distributive laws we express

$$((A \wedge B) \vee C)' = (A \wedge B)' \wedge C' = A' \wedge B' \wedge C'$$

Setting 1 for the value *True* and 0 for *False* for  $A, B, C$  we obtain the table for  $\varphi(A, B, C) = A' \wedge B' \wedge C'$ :

A	B	C	$\varphi$
0	0	0	1
0	0	1	0
0	1	0	1
1	0	0	1
0	1	1	0
1	0	1	0
1	1	0	0
1	1	1	0

□

**12.A.5.** For  $A = 0, B = 0, C = 1$  find the value of the Boolean proposition

$$A \wedge B \vee C$$

**Solution.**  $A \wedge B \vee C = 0 \wedge 0 \vee 1 = 0 \vee 1 = 1$ , which has value *True*. □

**12.A.6.** A switching circuit is an electrical network that consists of switches being open or closed as on fig. .. If switch  $a$  is open we set  $a = 0$ , if switch  $a$  is closed we set  $a = 1$  and so on. Switches labeled with the same letter must be either all open or all closed simultaneously. Switch  $a'$  is open if and only if switch  $a$  is closed. If current can flow between the extreme left and right ends of the circuit, the output of the circuit is 1 and 0 otherwise. The switching circuit representing  $a \wedge b$  is called series circuit, while the circuit representing  $a \vee b$  is called parallel circuit.

Represent the switching circuit at the figure?? as a Boolean proposition.

**Solution.** We can write this circuit as proposition  $\varphi = a \wedge X \wedge b$ , where

$$X = b \vee d' \vee c \wedge (a \vee d \vee c')$$

Thus,

$$\varphi = a \wedge (b \vee d' \vee c \wedge (a \vee d \vee c')) \wedge b.$$

□

**12.A.7.** Draw a switching circuit representing Boolean proposition

$$(a \wedge ((b \wedge c') \vee (b' \wedge c))) \vee (a \wedge b \wedge c)$$

**Solution.** see fig???

□

Properties (6)–(8) require the existence of neutral elements for both operations as well as the existence of an analog to the “inverse” of each element. (Note however, that the intersection with the complement results in the neutral element for union and vice versa. This is the other way round for numbers.)

There are only a few structures that possess such restrictive properties.

BOOLEAN ALGEBRAS

**Definition.** A set  $K$  together with two binary operations  $\wedge, \vee$ , a unary operation  $'$ , and two special elements  $1, 0$ , which satisfy the properties (1)–(8) is called a *Boolean algebra*. The  $\wedge$  operation is called *infimum* or *meet*, and the  $\vee$  operation is called *supremum* or *join*. The element  $A'$  is called the *complement* of  $A$ .

Note that the axioms of Boolean algebras are symmetric with respect to interchanging the operations  $\wedge$  and  $\vee$  together with the interchange of 0 and 1. This means that any proposition that can be derived from the axioms has a valid *dual proposition*, created by interchanging  $\wedge$  with  $\vee$  and 0 with 1. This is the *principle of duality*.

**12.1.3. Properties of Boolean algebras.** As usual, we derive several elementary corollaries of the axioms. In particular, note that in the special case of the Boolean algebra of all subsets of a given set, this proves all the elementary properties known from set algebra. The special elements 1 and 0 are unique as the neutral elements for  $\wedge$  and  $\vee$ . If there is  $\tilde{0}$  with the same properties then  $\tilde{0} = \tilde{0} \vee 0 = 0$ . Similarly 1 is also unique. Next, if  $B, C \in K$  satisfy the properties of  $A'$  (axiom (8) in the above definition), then

$$\begin{aligned} B &= B \vee 0 = B \vee (A \wedge C) = \\ &= (B \vee A) \wedge (B \vee C) = 1 \wedge (B \vee C) = B \vee C \end{aligned}$$

and similarly

$$C = C \vee B = B \vee C.$$

Therefore,  $B = C$  and so the complement to any  $A \in K$  is determined uniquely by its properties.

Generally the above observation means that given  $(K, \wedge, \vee)$ , there exists at most one operation  $'$  which completes it to a Boolean algebra  $(K, \wedge, \vee, ', 1, 0)$ , together with unique elements 1 and 0. Generally  $(K, \wedge, \vee)$  is written, omitting the other three symbols for operations.

The properties of the following proposition have their own names in set algebra: (2) is called *absorption laws*, (3) is the *idempotency* of the operations  $\wedge$  and  $\vee$ , and the equalities of (4) are the *De Morgan laws*.



**12.A.8.** Verify equivalence of the following two Boolean propositions

$$(a \vee b) \wedge (c \vee d) \quad \text{and} \quad (a \wedge c) \vee (a \wedge d) \vee (b \wedge c) \vee (b \wedge d)$$

**Solution.** Distributive laws yield

$$(a \vee b) \wedge (c \vee d) = ((a \vee b) \wedge c) \vee ((a \vee b) \wedge d) = a \wedge c \vee b \wedge c \vee a \wedge d \vee b \wedge d$$

□

**12.A.9.** Determine whether the following two Boolean propositions

$$((a' \wedge b) \vee (a \wedge c'))' \quad \text{and} \quad (a \vee b') \wedge (a \vee c')$$

are equivalent?

**Solution.** setting  $a = 1, b = 1, c = 0$  implies  $((a' \wedge b) \vee (a \wedge c'))' = 0$  and  $(a \vee b') \wedge (a \vee c') = 1$ . Hence, these propositions are not equivalent. □

**12.A.10.** Determine whether the following two Boolean propositions

$$(a \wedge b \wedge c) \vee (a \vee c)' \quad \text{and} \quad (a \wedge c) \vee (a' \wedge c')$$

are equivalent?

**Solution.** setting  $a = 1, b = 0, c = 1$  implies  $(a \wedge b \wedge c) \vee (a \vee c)' = 0$  and  $(a \wedge c) \vee (a' \wedge c') = 1$ . Hence, these propositions are not equivalent. □

**12.A.11.** Show that the negation  $a'$  can not be obtained by a combinatorial circuit that consists of  $\wedge$  and  $\vee$  operations only.

**Solution.** We prove the claim by the induction on the number of operations  $n$  in the circuit. For  $n = 1$  it is clear that neither proposition  $b \wedge c$  or  $b \vee c$  is equivalent to  $a'$ . Suppose  $a'$  can not be achieved by a combinatorial circuit, containing  $\vee$  and  $\wedge$ . Consider a circuit with  $n + 1$  operations and the first occurrence of the input of  $a$ . It can be either  $\wedge$  or  $\vee$ . If it is  $\wedge$  with the second argument either  $a$  or  $1$ , then the outcome of it is the same as input  $a$  and this operation  $\wedge$  can be deleted from the circuit with input  $a$  going to the next available operation. If the second argument is  $0$ , then  $a \wedge 0 = 0$  and operation can again be deleted with  $0$  going as input to the next available operation. If  $a$  enters the circuit at the  $\vee$ , and the second argument of this  $\vee$  is either  $a$  or  $0$ , then the  $\vee$  can be deleted with  $a$  entering the next operation. If the second argument equals  $1$ , then the result of  $\vee$  is  $1$ , and it can be entered to the next available operation. □

**12.A.12.** Simplify the formula

$$((A \wedge B) \vee (A \Rightarrow B)) \wedge ((B' \Rightarrow C) \vee (B \wedge C')).$$

**Solution.** In Boolean algebra, we obtain

$$(a \cdot b + a' + b) \cdot (b + c + b \cdot c') = \dots = a' \cdot c + b.$$

This means that the given formula is equivalent to  $(A' \wedge C) \vee B$ . □

FURTHER PROPERTIES OF BOOLEAN ALGEBRAS

**Proposition.** In every Boolean algebra  $(K, \wedge, \vee)$ ;

- (1)  $A \wedge 0 = 0, \quad A \vee 1 = 1,$
  - (2)  $A \wedge (A \vee B) = A, \quad A \vee (A \wedge B) = A,$
  - (3)  $A \wedge A = A, \quad A \vee A = A,$
  - (4)  $(A \wedge B)' = A' \vee B', \quad (A \vee B)' = A' \wedge B',$
  - (5)  $(A')' = A.$
- for all  $A, B \in K$ .

**PROOF.** By the principle of duality, it suffices to prove only one of the claims in each item. Begin with (3), of course using just the axioms of the Boolean algebras

$$A = A \wedge 1 = A \wedge (A \vee A') = (A \wedge A) \vee 0 = A \wedge A.$$

Now, (1) is proved easily:

$$A \wedge 0 = A \wedge (A \wedge A') = (A \wedge A) \wedge A' = A \wedge A' = 0.$$

(2) is also easy (read the second equality from right to left):

$$\begin{aligned} A \wedge (A \vee B) &= (A \vee 0) \wedge (A \vee B) = \\ &= A \vee (0 \wedge B) = A \vee 0 = A. \end{aligned}$$

In order to prove the De Morgan laws, it suffices to verify that  $A' \vee B'$  has the properties of the complement of  $A \wedge B$ . By the above, it must be the complement. Using (1), compute

$$\begin{aligned} (A \wedge B) \wedge (A' \vee B') &= ((A \wedge B) \wedge A') \vee ((A \wedge B) \wedge B') \\ &= (0 \wedge B) \vee (A \wedge 0) = 0. \end{aligned}$$

Similarly,

$$\begin{aligned} (A \wedge B) \vee (A' \vee B') &= (A \vee (A' \vee B')) \wedge (B \vee (A' \vee B')) \\ &= (1 \vee B') \wedge (1 \vee A') = 1. \end{aligned}$$

Finally, from the definition,  $A' \wedge A = 0$  and  $A' \vee A = 1$ . Hence,  $A$  has the required properties of the complement of  $A'$ , which means that  $A = (A')'$ . □

**12.1.4. Examples of Boolean algebras.** The intersection and union on all subsets in a given set  $M$  always define a Boolean algebra. The smallest is the set of all subsets of a singleton  $M$ . It contains two elements, namely  $0 = \emptyset$  and  $1 = M$  with the obvious equalities  $0 \wedge 1 = 0, 0 \vee 1 = 1$ , etc. The operations  $\wedge$  and  $\vee$  are the same as multiplication and addition in the remainder class ring  $\mathbb{Z}_2$  of even and odd numbers. This is called the Boolean algebra  $\mathbb{Z}_2$ . This is the only case when a Boolean algebra is a field of scalars at the same time!



As in the case of rings of scalars or vector spaces, the algebraic structure of a Boolean algebra can be extended to all spaces of functions whose codomain is a Boolean algebra. For the set  $S = \{f : M \rightarrow K\}$  of all functions from a set  $M$  to a Boolean algebra  $(K, \wedge, \vee)$ , the necessary operations and the distinguished elements  $0$  and  $1$  on  $S$  can be defined



**12.A.13.** Design a logic proposition in three boolean that takes value *True* if and only if the majority of variables also have *True* values.

**Solution.** For this purpose, a dijunctive normal form is quite suitable. Consider the proposition:

$$a \wedge b \wedge c' \vee a \wedge b' \wedge c \vee a' \wedge b \wedge c \vee a \wedge b \wedge c.$$

**12.A.14.** Let  $A$  be a free boolean algebra, generated by countably many generators  $a_1, \dots, a_n, \dots$  with operations  $\wedge, \vee$  and  $'$  and elements being finitely many of these generators, linked by the operations. Prove that these algebra is atomless.

**Solution.** Suppose  $f$  is an atom in  $A$ . Without loss of generality  $f$  can be expressed in terms of the first  $n$  generators  $a_1, \dots, a_n$ . Then  $f \wedge a_{n+1}$  is neither  $f$  nor  $0$ , so  $f$  can not be an atom in  $A$ .  $\square$

**12.A.15.** Prove monotonicity laws for Boolean algebras: if  $A \leq \tilde{A}$  and  $B \leq \tilde{B}$  then

- i)  $A \vee B \leq \tilde{A} \vee \tilde{B}$ ;
- ii)  $A \wedge B \leq \tilde{A} \wedge \tilde{B}$ ;
- iii)  $(\tilde{A})' \leq A'$

**Solution.** Since

$$(A \vee B) \vee (\tilde{A} \vee \tilde{B}) = (A \vee \tilde{A}) \vee (B \vee \tilde{B}) = \tilde{A} \vee \tilde{B},$$

it implies the first assertion. The second follows from the duality the fact that  $A \leq \tilde{A}$  if and only if  $A \wedge \tilde{A} = A$ . Application of de Morgan's law

$$(\tilde{A})' \vee A' = (A \wedge \tilde{A}') = A'$$

verifies the third.  $\square$

**12.A.16.** For Boolean algebras prove

- i)  $A \leq B$  if and only if  $B' \leq A'$ ;
- ii)  $A \leq B$  if and only if  $A \wedge B' = 0$ ;
- iii)  $C \wedge A \leq B$  if and only if  $A \leq C' \vee B$ .

**Solution.** By previous exercise,  $A \leq B$  yields  $B' \leq A'$ , and vice versa

$$A = (A')' \leq (B')' = B.$$

If  $A \leq B$ , then  $A = A \wedge B$  and  $A \wedge B' = A \wedge B \wedge B' = 0$ .

Conversely, from  $A \wedge B' = 0$  follows that

$$A = A \wedge (B \vee B') = A \wedge B \vee A \wedge B' = A \wedge B,$$

so  $A \leq B$ . For the third assertion, since  $C \wedge A \leq B$  and  $C' \wedge A \leq C'$ , by monotonicity of the  $\vee$ ,

$$A = C \wedge A \vee C' \wedge A \leq B \vee C'.$$

On the other hand, if  $A \leq C' \vee B$ , by monotonicity of the  $\wedge$ ,

$$C \wedge A \leq C \wedge (C' \vee B) \leq C \wedge B \leq B.$$

as functions of an argument  $x \in M$  as follows:

$$(f_1 \wedge f_2)(x) = (f_1(x)) \wedge (f_2(x)) \in K,$$

$$(f_1 \vee f_2)(x) = (f_1(x)) \vee (f_2(x)) \in K,$$

$$(1)(x) = 1 \in K, (0)(x) = 0 \in K,$$

$$(f)')(x) = (f(x))' \in K.$$

$\square$  It is easy and straightforward to verify that these new operations define a Boolean algebra.

Recall that the subsets of a given set  $M$  can be viewed as mappings  $M \rightarrow \mathbb{Z}_2$  (the elements of the subset in question are mapped to 1 while all others go to 0). Then, the union and intersection can be defined in the above manner — for instance, evaluating the expression  $(A \vee B)(x)$  for a point  $x \in M$ , This determines whether it lies in  $A$  or whether it lies in  $B$ , and whether the join of the results is in  $\mathbb{Z}_2$ . The result is 1 if and only if  $x$  lies in the union.  $\square$

**12.1.5. Propositional logic.** The latter simple observation



brings us close to the calculus of elementary logic. View the notations for operations in a Boolean algebra as creating “words” from the elements  $A, B, \dots \in K$ , the operations  $\vee, \wedge, '$  and parentheses, which clarify the desired precedence of the operations.

The axioms of the Boolean algebras and their corollaries say how different words may produce the same result in  $K$ .

This is clear in the case of  $K = 2^M$ , the set of all subsets of a given set; it is just equality of subsets. Now, another interpretation is mentioned in terms of operations in formal logic.

Work with words as above but view them as propositions composed from elementary (atomic) propositions  $A, B, \dots$  and the logical operations AND (the binary operation  $\wedge$ ), OR (the binary operation  $\vee$ ), and the negation NOT (the unary operation  $'$ ). These words are called *propositions*. They are assigned a truth value depending on the truth values of the individual atomic propositions. The truth value is an element of the trivial Boolean algebra  $\mathbb{Z}_2$ , i.e. either 0 or 1.

The truth value of a proposition is completely determined by assigning the truth values for the simplest propositions  $A \wedge B, A \vee B$  and  $A'$ .  $A \wedge B$  is defined to be true if and only if both  $A$  and  $B$  are true.  $A \vee B$  is false if and only if both  $A$  and  $B$  are false. The value of  $A'$  is complementary to  $A$ .

A proposition with  $n$  elementary propositions defines a function  $(\mathbb{Z}_2)^n \rightarrow \mathbb{Z}_2$ . Two propositions are called logically equivalent if and only if they define the same function. In the previous paragraph, it is already verified that the set of all classes of logically equivalent propositions has the structure of a Boolean algebra. Propositional logic satisfies everything proved for general Boolean algebras.

Next, we consider how other usual simple propositions of propositional logic are represented as elements of the Boolean algebra. Expressions always correspond to a class of logically equivalent propositions:  $\square$

**12.A.17.** For any two elements  $A, B$  of a Boolean algebra, the *symmetric difference* of  $A$  and  $B$  is defined as

$$A \triangle B = A \wedge B' \vee B \wedge A'.$$

This is analogous of taking symmetric difference  $A \triangle B = (A \setminus B) \cup (B \setminus A)$ . Show that

- i)  $A = B$  if and only if  $A \triangle B = 0$ .
- ii)  $A \triangle (B \triangle C) = (A \triangle B) \triangle C$
- iii)  $A \wedge (B \triangle C) = (A \wedge B) \triangle (A \wedge C)$ .

**Solution.**

- i) If  $A = B$  then  $A \leq B$  and  $B \leq A$ , and therefore, by previous exercise,  $A \wedge B' = 0$  and  $B \wedge A' = 0$ , thus  $A \triangle B = 0$ . On the other hand,  $A \triangle B = 0$  implies that both  $A \wedge B' = 0$  and  $B \wedge A' = 0$ , which yield  $A \leq B$  and  $B \leq A$ , thus  $A = B$ .
- ii) Expanding  $(B \triangle C)'$  using distributivity and de Morgan's laws, we obtain

$$(B \triangle C)' = (B' \vee C) \wedge (C' \vee B) = B \wedge C \vee B' \wedge C'.$$

Further expanding  $A \triangle (B \triangle C)$  we get

$$\begin{aligned} A \triangle (B \triangle C) &= A \wedge (B \wedge C \vee B' \wedge C') \vee A' \wedge (B \wedge C' \vee B' \wedge C) \\ &= A \wedge B \wedge C \vee A \wedge B' \wedge C' \vee A' \wedge B \wedge C' \vee A' \wedge B' \wedge C, \end{aligned}$$

which equals  $(A \triangle B) \triangle C$  by obvious symmetry.

iii) Follows from

$$A \wedge (B \triangle C) = A \wedge (B \wedge C' \vee C \wedge B') = A \wedge B \wedge C' \vee A \wedge C \wedge B'$$

and

$$(A \wedge B) \triangle (A \wedge C) = A \wedge B \wedge (A' \vee C') \vee A \wedge C \wedge (A' \vee B') = A \wedge B \wedge C' \vee A \wedge C \wedge B'$$

□

**12.A.18.** The Sheffer stroke is defined as:

$$A | B = A' \wedge B'.$$

The *Pierce arrow* can be defined as:

$$A \uparrow B = A' \vee B'.$$

Prove that  $|$  and  $\uparrow$  define  $A \vee B$ ,  $A \wedge B$  and  $A'$ .

**Solution.** It is clear from de Morgan's laws, in particular, that  $A' = A \uparrow A$ ,  $A \vee B = (A \uparrow A) \uparrow (B \uparrow B)$ . Since  $A \wedge B = (A' \vee B')'$ , we obtain that  $\uparrow$  generates all Boolean operations. Analogously,  $A' = A | A$  and  $A \wedge B = A' | B' = (A | A) | (B | B)$ . De Morgan's law  $A \vee B = (A' \wedge B')'$  proves that  $|$  also generates all three Boolean operations. □

**12.A.19.** The *exclusive or* binary operation in Boolean algebra is defined as:

$$A \oplus B = (A \vee B) \wedge (A' \vee B').$$

Prove that  $\oplus$  and  $\wedge$  do not generate all Boolean operations.

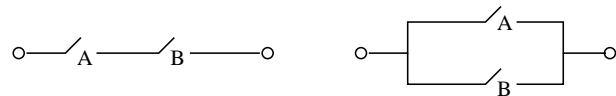
**Solution.** If we only apply  $\oplus$  and  $\wedge$  operations, from the inputs with all values *False* we'll be getting outputs with the same *False* value. Thus,  $A'$  can not be obtained using only  $\oplus$  and  $\wedge$  out of a *False* proposition  $A$ . □

THE STANDARD LOGICAL OPERATORS

- (1)  $A$  AND  $B$  corresponds to  $A \wedge B$ ,
- (2)  $A$  OR  $B$  corresponds to  $A \vee B$ ,
- (3) the implication  $A \Rightarrow B$  can be obtained as  $A' \vee B$ ,
- (4) the equivalence  $A \Leftrightarrow B$  corresponds to  $(A \wedge B) \vee (A' \wedge B')$ ,
- (5) the exclusive OR, known as  $A$  XOR  $B$ , is given as  $(A \wedge B') \vee (A' \wedge B)$ ,
- (6) the negation of OR,  $A$  NOR  $B$ , is expressed as  $A' \wedge B'$ ,
- (7) the negation of AND,  $A$  NAND  $B$ , is given as  $A' \vee B'$ ,
- (8) tautology (proposition always true) is given in terms of an arbitrary atomic proposition as  $A \vee A'$ , its negation (always false) is  $A \wedge A'$ .

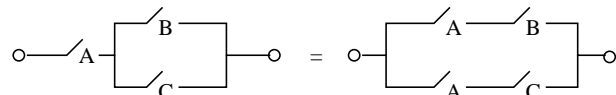
Note that in set algebra, XOR corresponds to the symmetric difference.

**12.1.6. Switch boards as Boolean algebras.** A switch is a black box with only two states – it is either on (and the signal goes through) or off (and the signal does not go through).



One or more switches may be interconnected in a series circuit or a parallel circuit. The series circuit corresponds to the binary operation  $\wedge$ , while the parallel circuit corresponds to  $\vee$ . The unary operation  $A'$  defines a switch whose state is always the opposite than that of  $A$ . Every finite word created from the switches  $A, B, \dots$  and the operations  $\wedge, \vee$ , and  $'$  can be transformed to a diagram that represents a system of switches, connected by wires, similarly as in the above subsection, where each choice of states of the individual switches gives the value “on/off” for the entire system. The discussion is about switchboards and their logical evaluation function.

Again, it is easy to verify all the axioms of Boolean algebras for the system. The following diagram illustrates one of the distributivity axioms.



The circuit without a switch corresponds to 1. When the endpoints are not connected, this corresponds to 0 (consider a series circuit of  $A$  and  $A'$ ). Draw diagrams for all axioms of Boolean algebras and verify them!

We return to this example shortly, showing that each expression in propositional logic can be modeled by a switch board.

**12.1.7. Algebra of divisors.** There are other natural examples of Boolean algebras. Choose a positive integer  $p \in \mathbb{N}$ . The underlying set  $D_p$  is the set of all divisors  $q$  of  $p$ . For two such divisors  $q, r$ , define  $q \wedge r$  to be the greatest common divisor of  $q$  and  $r$ , and  $q \vee r$  is defined to be their least common multiple (cf. the



**12.A.20.** Translate the following sentence as a logical proposition considering ambiguity of English language: *Either school X improves its performance and continues to perform well or school Y wins academic competition.*

**Solution.** Modern English is sometimes vague about the meaning of *or*. Therefore, the sentence can be translated using  $\vee$  or  $\oplus$  in place of the *or*. Let  $A, B, C$  represent propositions *School X improves its performance*, *School X continues to perform well*, and *School Y wins academic competition*. Then the proposition of the sentence can be interpreted as

- i)  $(A \wedge B) \vee C$ ;
- ii)  $(A \wedge B) \oplus C$ ;

This two propositions differ when all three propositions  $A, B, C$  have *True* value. □

**12.A.21.** Let logical implication  $\rightarrow$  mean  $A \rightarrow B = A' \vee B$ . Prove the equivalence of the following logical propositions

- i)  $(A' \rightarrow (B \rightarrow C)) = (B \rightarrow (A \vee C))$ ;
- ii)  $(A \rightarrow B')' \wedge (A \vee B)' = 0$ , that is, it is always *False*.
- iii)  $(A \vee B) \wedge (B' \vee C) \rightarrow (B \vee C) = 1$  that is, it is always *True*

**12.A.22.** Show that the following divisibility relations are partial orders on  $X$ . Is any one of them a linear order?

- i)  $X = \mathbb{N}$ . Relation  $|$  is defined as:  $m|n$  if  $m$  divides  $n$ .
- ii) Let  $X$  be a set of all integer divisors of 36.

**Solution.**

- i) As  $n|n$  for each  $n$ , the relation is reflexive. IF  $m|n$  and  $n|m$ , then  $m = n$ . Since  $r|m$  and  $m|n$  implies  $r|n$ , it is also transitive. The order is not linear, since, for example, 4 and 5 are not divisors of each other.
- ii)  $X = \{1, 2, 3, 4, 6, 9, 12, 18, 36\}$ . By previous item,  $|$  is reflexive, antisymmetric and transitive. However, divisors 2 and 3 are not related. (DRAW HASSE DIAGRAM!) □

**12.A.23.** Given a partial order  $\leq$  on a set  $A$ , one can also define a corresponding strict order  $<$ , that is also useful in various situations. A relation  $R$  on  $A$  is called irreflexive if  $(x, x) \notin R$  for all  $x \in A$ .  $R$  is called asymmetric if with every  $(x, y) \in R$ ,  $(y, x) \notin R$ . Relation  $R$  on  $A$  is a strict order if  $R$  is a partial order and, in addition, it is irreflexive and asymmetric. Let  $A$  be a set

- i) If  $\leq$  is a partial order on  $A$ , then the relation  $<$ , defined as  $a < b$  if  $a \leq b$  and  $a \neq b$  is a strict order on  $A$ .
- ii) Given strict order  $<$  on  $A$  then the partial order  $\leq$  can be defined as  $a \leq b$  if  $a < b$  or  $a = b$ .

**Solution.** Suppose,  $\leq$  is a partial order with  $<$  defined by  $a < b$  for  $a \leq b$  and  $a \neq b$ . Then  $<$  is irreflexive: if  $a < b$  then  $a \neq b$ . Now, suppose that  $a < b$  and  $b < c$ , then  $a \leq c$  by transitivity of  $\leq$ . If  $a = c$ , then by asymmetry of  $\leq$ , we would have  $a = b$ . Hence,  $a < c$ . □

previous chapter for the definitions and context). The distinguished element  $1 \in D_p$  is defined to be  $p$  itself. The neutral element 0 for join on  $D_p$  is the integer  $1 \in \mathbb{N}$ . The unary operation  $'$  is defined using division as  $q' = p/q$ .

**Proposition.** *The set  $D_p$  together with the above operations  $\wedge, \vee$ , and  $'$  is a Boolean algebra if and only if the factorization of  $p$  contains no squares (i.e., in the unique factorization  $p = q_1 \dots q_n$ , the prime numbers  $q_i$  are pairwise distinct).*

**PROOF.** It is easy to verify the axioms of Boolean algebras under the assumptions of the proposition. It might be interesting to see where the assumption squarefree is needed.

The greatest common divisor of a finite number of integers is independent of their order. This also holds for the least common multiple. This corresponds to the axioms (1) and (2) in 12.1.2. The commutativity (3) is clear.

For any three elements  $a, b, c$ , write their factorizations without loss of generality as  $a = q_1^{n_1} \dots q_s^{n_s}$ ,  $b = q_1^{m_1} \dots q_s^{m_s}$ ,  $c = q_1^{k_1} \dots q_s^{k_s}$ . Zero powers are allowed and all  $q_j$  are pairwise coprime. Thus,  $a \wedge b \in D_p$  corresponds to the element in which each  $q_i$  occurs with the power that is the minimum of the powers in  $a$  and  $b$ . This holds analogously for  $a \vee b$  and maximum. The distributivity laws (4) and (5) of 12.1.2 now follow easily.

There is no problem with the existence of the distinguished elements 0 and 1. These are already defined directly and clearly satisfy the axioms (6) and (7). However, if there are squares in the factorizations, then this prevents the existence of complements. For instance, in  $D_{12} = \{1, 2, 3, 4, 6, 12\}$ ,  $6 \wedge 6' = 1$  cannot be reached since 6 has a non-trivial divisor which is common to all other elements of  $D_{12}$  except for 1, but  $6 \vee 1 = 6 \neq 12$ . (The number 1 is the potential smallest element in  $D_{12}$ ; it plays the role of 0 from the axioms.)

Nevertheless, if there are no squares in the factorization of the integer  $p$ , then the complement can be defined as  $q' = p/q$ , and it can be verified easily that this definition satisfies the axiom 12.1.2(8). □

If there are no squares in the decomposition of  $p$ , then the number of all divisors is a power of 2. This suggests that these Boolean algebras are very similar to the set algebras we started with. We return to the classification of all finite Boolean algebras. Before that, we consider structures like the divisors above for general  $p$ .

**12.1.8. Partial order.** There is a much more fundamental concept, the *partial order*. See the end of chapter 1. Recall that the definition of a partial order is a reflexive, antisymmetric, and transitive relation  $\leq$  on a set  $K$ . A set with partial order  $(K, \leq)$  is called a *partially ordered set*, or *poset* for short. The adjective “partial” means that in general, this relation does not say whether  $a \leq b$  or  $b \leq a$  for every two different elements  $a, b \in K$ . If it does for each pair, it is called a *linear order* or a *total order*.



**12.A.24.** Prove that a finite poset  $A$  can be embedded in a totally ordered set, whose order extends the original order on  $A$ .

**Solution.** Let  $(A, \leq)$  be a finite poset with a minimal element  $a_1$ . Set  $A \setminus \{a_1\}$  remains a poset with some minimal element  $a_2$ . Choose  $a_3, \dots, a_n$  similarly until  $A$  is exhausted. Define  $a_i \leq a_j$  for  $i \leq j$ , which defines a total order on  $A$ .  $\square$

**12.A.25.** Convert the poset of divisors of 36 into a linearly ordered set in different ways

**Solution.** One order is the natural order in  $\mathbb{N}$ : 1, 2, 3, 4, 6, 9, 12, 18, 36. Another order can be chosen as 1, 2, 4, 3, 6, 12, 9, 18, 36.  $\square$

**12.A.26.** Prove that lattices of the *diamond* and *pentagon* types are not distributive.

**Solution.**

i) For the diamond lattice

$$a \wedge (b \vee c) = a \wedge 1 = a, \quad \text{while} \quad (a \wedge b) \vee (a \wedge c) = 0 \vee 0 = 0.$$

ii) For the pentagon lattice

$$a \wedge (b \vee c) = a \wedge 1 = a, \quad \text{and} \quad (a \wedge b) \vee (a \wedge c) = 0 \vee 0 = 0.$$

also hold.  $\square$

**12.A.27.** Prove uniqueness of completeness in distributive lattices, i.e. if  $(A, \leq)$  is a bounded distributive lattice with minimum 0 and maximum 1 then complements are unique.

**Solution.** Let  $x'$  and  $z$  denote complements of  $x \in A$ . Then

$$\begin{aligned} (1) \quad x' &= x' \wedge 1 \\ &= x' \wedge (x \vee z) \\ &= (x' \wedge x) \vee (x' \wedge z) \\ &= 0 \vee (x' \wedge z) \\ &= (x' \wedge z) \vee (x' \wedge z) \\ &= (x \vee x') \wedge z \\ &= 1 \wedge z = z \end{aligned} \quad \square$$

**12.A.28.** Anne, Brenda, Kate, and Dana want to set out on a trip. Find out which of the girls will go if the following must hold: At least one of Brenda and Dana will go; at most one of Anne and Kate will go; at least one of Anne and Dana will go; at most one of Brenda and Kate will go; Brenda will not go unless Anne goes; and Kate will go if Dana goes.

**Solution.** Transforming the problem to Boolean algebra, simplifying it, and transforming it back, we find out that either Anne and Brenda will go or Kate and Dana will go.  $\square$

There is always a partial order on the set  $K = 2^M$  of all subsets of a given set  $M$  – the inclusion. In terms of intersections or joins, the inclusion can be defined as  $A \subseteq B$  if and only if  $A \wedge B = A$ , or equivalently,  $A \subseteq B$  if and only if  $A \vee B = B$ . In general, each Boolean algebra is a very special poset:

**Lemma.** Let  $(K, \wedge, \vee)$  be a Boolean algebra. Then the relation  $\leq$  defined by  $A \leq B$  if and only if  $A \wedge B = A$  is a partial order. Moreover, for all  $A, B, C \in K$ :

- (1)  $A \wedge B \leq A$ ,
- (2)  $A \leq A \vee B$ ,
- (3) if  $A \leq C$  and  $B \leq C$ , then  $A \vee B \leq C$ ,
- (4)  $A \leq B$  if and only if  $A \wedge B' = 0$ ,
- (5)  $0 \leq A$  and  $A \leq 1$ .

**PROOF.** All the properties to be proved are results of simple calculations in the Boolean algebra  $K$ . Begin with the properties of a partial order for  $\leq$ . Reflexivity is a direct corollary of idempotency:  $A \wedge A = A$ , i.e.  $A \leq A$ . Similarly, the commutativity of  $\wedge$  guarantees the antisymmetry of  $\leq$ , since if both  $A \wedge B = A$  and  $B \wedge A = B$ , then

$$A = A \wedge B = B \wedge A = B.$$

Finally, if  $A \wedge B = A$  and  $B \wedge C = B$ , then

$$A \wedge C = (A \wedge B) \wedge C = A \wedge (B \wedge C) = A \wedge B = A,$$

which verifies the transitivity of  $\leq$ .

Similarly,  $(A \wedge B) \wedge A = (A \wedge A) \wedge B = A \wedge B$ , that is,  $A \wedge B \leq A$ .

It follows from  $A \wedge (A \vee B) = A$ , see 12.1.3(2), that  $A \leq A \vee B$ . This proves the claim (2).

Distributivity together with assumption (3) provides

$$(A \vee B) \wedge C = (A \wedge C) \vee (B \wedge C) = A \vee B,$$

so that (3) holds.

The proposition (5) follows directly from the axioms for the distinguished elements 1 and 0.

It remains to prove (4). If  $A \leq B$ , then  $A \wedge B' = A \wedge B \wedge B' = 0$ . On the other hand, if  $A \wedge B' = 0$ , then  $A = A \wedge 1 = A \wedge (B \vee B') = (A \wedge B) \vee (A \wedge B') = (A \wedge B) \vee 0 = A \wedge B$ . Hence  $A \leq B$ , and the proof is finished.  $\square$

Note that as for the algebra of subsets, in all Boolean algebras  $A \wedge B = A$  if and only if  $A \vee B = B$ . If  $A \wedge B = A$ , then the absorption laws imply that  $A \vee B = (A \wedge B) \vee B = B$ , and vice versa. Therefore, the operation  $\vee$  can also be used in the definition of a partial order.

Every poset  $(K, \leq)$  corresponds to a (oriented) graph (cf. the beginning of chapter 13 for definitions if necessary): the vertex set is  $K$ , and there is an edge leading from  $a$  to  $b$  if and only if  $a \leq b$ . This is a convenient way how to represent finite posets.

A *Hasse diagram* of a poset is a drawing of this graph in the plane so that greater elements are drawn above lower ones. Since the edge orientation is implicitly given by this, it need not be drawn explicitly. Furthermore, loops and edges

**12.A.29.** Solve the following problem by transforming it to Boolean algebra: Tom, Paul, Sam, Ralph, and Lucas are suspected of having committed a murder. It is certain that at the crime scene, there were: at least one of Tom and Ralph, at most one of Lucas and Paul, and at least one of Lucas and Tom. Sam could be there only if so was Ralph. However, if Sam was there, then so was Tom. Paul could never cooperate with Ralph, but Paul and Tom are an inseparable pair. Who committed the murder?



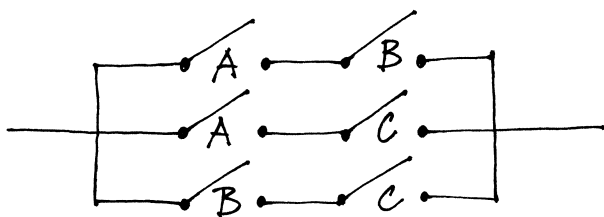
**Solution.** Transforming into Boolean algebra, using the first letter of each name, we get

$$(t + r)(l' + p')(l + t)(r + s')(s' + t)(p' + r')(pt + p't')$$

and thanks to  $x^2 = x$ ,  $xx' = 0$ ,  $x + x' = 1$ , we can rearrange the above to  $s'r'ptl' + s'rp't'l$ . Thus, the murder was committed either by Tom and Paul or by Ralph and Lucas.  $\square$

**12.A.30.** A vote box for three voters is a box which processes three votes and outputs “yes” if and only if majority of the voters is for. Design this box using from switch circuits.

**Solution.**



$\square$

**12.A.31.** Find a finite subset of the set of positive integers which is not a lattice with respect to divisibility.  $\circ$

**12.A.32.** Find the number of partial orders on a given 4-element set. Draw the Hasse diagram of each isomorphism class and determine whether it is a lattice. Is one of them a Boolean algebra?

**Solution.** We go through all Hasse diagrams of the partial orders on a 4-element set  $M$  and for each diagram, we count the number of partial orders (i. e. subsets of  $M \times M$ ) that correspond to it; see the picture:

$\dots$	$\vdots$	$\vdots \vdots$	$\wedge$	$\vee$	$\vdots$	$N$	$\times$
1	$\frac{4!}{2}$	$\frac{4!}{2}$	$\frac{4!}{2}$	$\frac{4!}{2}$	$4!$	$4!$	$\frac{4!}{4}$
$\wedge$	$\vee$	$\vdots$	$\vdots$	$\vee$	$\wedge$	$\diamond$	$\vdots$
$\frac{4!}{2}$	$\frac{4!}{6}$	$4!$	$4!$	$\frac{4!}{2}$	$\frac{4!}{2}$	$\frac{4!}{2}$	$4!$

Therefore, there are 219 partial orders on a given 4-element set.

Note that the condition of existence of suprema and infima of any pair of elements in a lattice implies (by induction) the existence of them for any finite non-empty subset. In particular, this means that every non-empty finite lattice has a greatest element as well as a least element.

which are implied by transitivity and reflexivity are omitted in the diagram. Especially when  $K$  has only a few elements, this is a very transparent way of discussing several cases; see the examples in the exercise column.

**12.1.9. Lattices.** Not every poset is created the latter way from a Boolean algebra. For instance, the trivial partial order is defined on any set as  $A \leq A$  for each  $A$ , but all pairs of different elements are incomparable. Such a poset cannot rise from a Boolean algebra if  $K$  contains more than one element (as seen, the least and greatest elements of a Boolean algebra are comparable to every element).



Think to what extent the operations  $\wedge$  and  $\vee$  can be built from a partial order. They are the suprema and infima in the following definition:

LOWER AND UPPER BOUNDS, SUPREMA, INFIMA

Consider a fixed poset  $(K, \leq)$ . An element  $C \in K$  is said to be a *lower bound* for a subset  $L \subseteq K$  if and only if  $C \leq A$  for all  $A \in L$ . An element  $C \in K$  is said to be the *greatest lower bound* (or *infimum*) of a subset  $L \subseteq K$  if and only if it is a lower bound and for every lower bound  $D$  of  $L$ ,  $D \leq C$ .

By replacing  $\leq$  with  $\geq$  in the above, the definitions of an *upper bound* and of the *least upper bound* (or *supremum*) of a subset  $L$  are obtained.

If the suprema and infima exist for all couples  $A, B$ , they define the binary operations  $\vee$  and  $\wedge$ , respectively.

LATTICES

**Definition.** A *lattice* is a poset  $(K, \leq)$  where every two-element set  $\{A, B\}$  has a supremum  $A \vee B$  and an infimum  $A \wedge B$ .

The poset  $(K, \leq)$  is said to be a *complete lattice* if and only if every subset of  $K$  has a supremum and an infimum.

The binary operations  $\wedge$  and  $\vee$  on a lattice  $(K, \leq)$  are clearly commutative and associative (prove this in detail!). The latter properties (of associativity and commutativity) ensure that all finite non-empty subsets in  $K$  possess infima and suprema.

Note that any element of a lattice  $K$  is an upper bound for the empty set. Thus in a complete lattice, the supremum of the empty set is the least element  $0$  of  $K$ . Similarly, the infimum of the empty set is the greatest element  $1$  of  $K$ . Of course, a finite lattice  $(K, \leq)$  is always complete (with  $1$  being the supremum of all elements in  $K$  and  $0$  the infimum of all elements in  $K$ ).

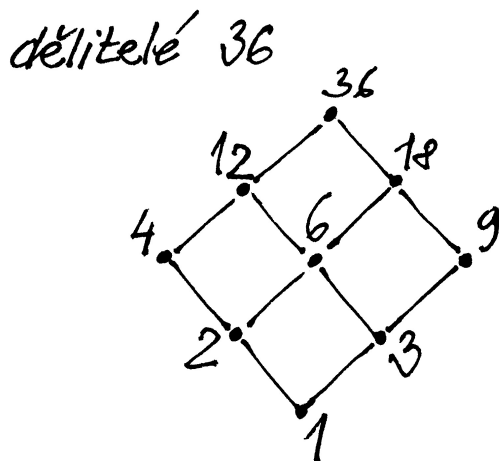
**Remark.** The poset of real numbers, completed with the greatest and least elements  $\pm\infty$ , is a complete lattice (with the standard ordering). We may view it as a completion of the poset of rational numbers. The classical Dedekind cut construction of this completion of rational numbers extends

Using this criterion, we can see that only the last two Hasse diagrams may be lattices. Indeed, they are lattices; the first one is even a Boolean algebra.  $\square$

**12.A.33.** Find the number of partial orders on the set  $\{1, 2, 3, 4, 5\}$  such that there are exactly two pairs of incomparable elements.  $\circ$

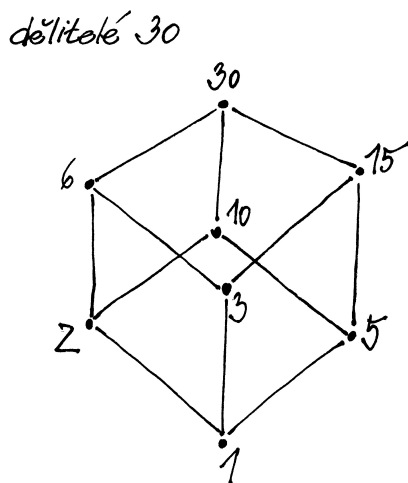
**12.A.34.** Draw the Hasse diagram of the lattice of all (positive) divisors 36. Is this lattice distributive? Is it a Boolean algebra?

**Solution.** The lattice is distributive (it does not contain a sublattice isomorphic to diamond or pentagon).



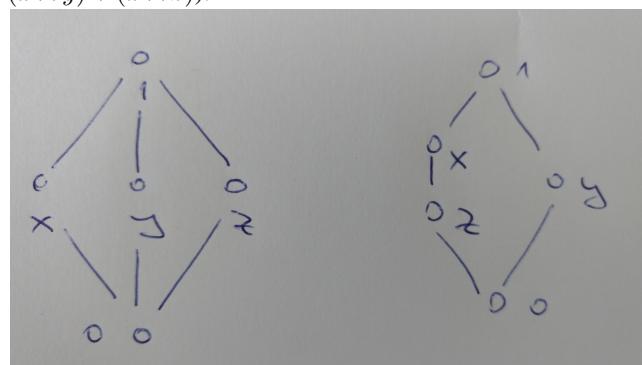
**12.A.35.** Draw the Hasse diagram of the lattice of all (positive) divisors 30. Is this lattice distributive? Is it a Boolean algebra?

**Solution.** This lattice is a Boolean algebra, and it has 8 elements. All finite Boolean algebras are of size  $2^n$  for an appropriate  $n$ , and they are all isomorphic for a fixed  $n$  (see 12.1.16). This Boolean algebra is a “cube”: its graph can be drawn as projection of a cube onto the plane.



to all posets. Indeed, the so called Dedekind-MacNeille completion provides the unique smallest complete lattice containing a given poset. We shall not go into any details here.

A lattice is said to be *distributive* if and only the operations  $\wedge$  and  $\vee$  satisfy the distributivity axioms (4) and (5) of subsection 12.1.2 on page 1041. There are lattices which are not distributive; see the Hasse diagrams of two such simple lattices below (and check that in both cases  $x \wedge (y \vee z) \neq (x \wedge y) \vee (x \wedge z)$ ).



Now, Boolean algebras can be defined in terms of lattices: a Boolean algebra is a complete distributive lattice such that each element has its complement (i.e. the axiom 12.1.2(8) is satisfied).

It is already verified that the latter requirement implies that complements are unique (see the ideas at the beginning of subsection 12.1.3), which means that the alternative definition of Boolean algebras is correct.

During the discussion of divisors of any given integer  $p$ , distributive lattices  $D_p$  are encountered. These distributive lattices are Boolean algebras if and only if  $p$  is squarefree, see 12.1.7.

**12.1.10. Homomorphisms.** Dealing with mathematical structures, most information about objects can be obtained/understood from the homomorphisms. These are mappings which preserve the corresponding operations. The linear mappings between vector spaces, or continuous mappings on  $\mathbb{R}^n$  or any metric spaces with the given topology of open neighbourhoods represent very good examples.

This concept is particularly simple for posets:

POSET HOMOMORPHISMS

Let  $(K, \leq_K)$  and  $(L, \leq_L)$  be posets. A mapping  $f : K \rightarrow L$  is called a *poset homomorphism* (also *order-preserving mapping*, *monotone mapping* or *isotone mapping*) if for all  $A \leq_K B$  the same relation  $f(A) \leq_L f(B)$  is true.

Although the structure of any Boolean algebra is completely determined by its subordinated poset structure, an isotone mapping does not necessarily respect the suprema and infima. Non-comparable elements  $A, B$  can be mapped to the same image  $f(A) = f(B)$ , while their suprema could map to  $f(A \vee B)$  strictly larger.

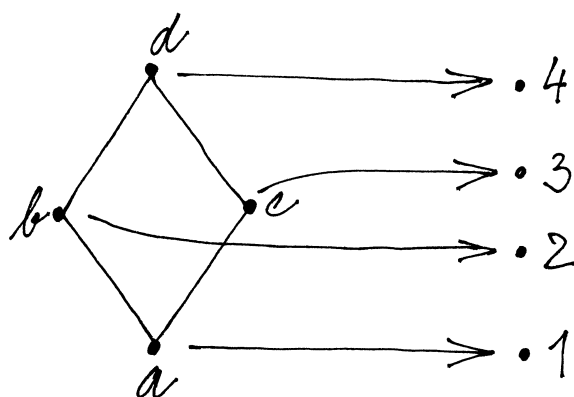
add a simple example to the other column, e.g. complete the graf K3,3 as Hasse diagram of a poset.

**12.A.36.** Decide whether every lattice on a 3-element set is a *chain*, i. e., whether each pair of elements are necessary comparable.

**Solution.** As we have noticed in exercise 12.A.32, every finite non-empty lattice must contain a greatest element and a least element. Each of these is thus comparable to any other, which means that the remaining one is comparable with these two; and there are no other elements.  $\square$

**12.A.37.** Find an example of two lattices and a poset homomorphism between them which is not a lattice homomorphism.

**Solution.** Again, we return to exercise 12.A.32 and consider the following mapping:



**12.A.38.** Decide whether every lattice homomorphism between finite non-empty lattices  $K, L$  maps the least element of  $K$  to the least element of  $L$ .

**Solution.** No, any constant mapping between two lattices is a lattice homomorphism. Thus, if we sent everything to an element different from the least one, we get the wanted counterexample homomorphism.  $\square$

**12.A.39.** Decide whether every chain which has a greatest element and a least element is a complete lattice.

**Solution.** No. Consider the set of non-zero integers and order it as follows: any positive integer is greater than any negative integer, but the ordering among the positive integers will be reversed, as well as among the negative integers. Then, 1 will be the greatest element of the resulting chain, and  $-1$  will be the least element. However, the subset of all positive integers does not have an infimum in this poset.

Formally we define the linear order  $\prec$  on  $\mathbb{Z} \setminus \{0\}$  by:

$$a \prec b \iff [(\text{sgn}(a) \cdot \text{sgn}(b) = 1 \wedge b > a) \vee (\text{sgn}(a) > \text{sgn}(b))].$$

$\square$  In the case of Boolean algebras, homomorphisms are defined as follows:

LATTICE AND BOOLEAN-ALGEBRA HOMOMORPHISMS

A mapping  $f : (K, \wedge, \vee) \rightarrow (L, \wedge, \vee)$  is a *homomorphism of Boolean algebras* if and only if for all  $A, B \in K$

- (1)  $f(A \wedge B) = f(A) \wedge f(B)$ ,
- (2)  $f(A \vee B) = f(A) \vee f(B)$ ,
- (3)  $f(A') = f(A)'$ .

Moreover, if  $f$  is bijective, it is an *isomorphism of Boolean algebras*.

Similarly, lattice homomorphisms are defined as mappings which satisfy the properties (1) and (2).

It is easily verified that if a homomorphism  $f$  is bijective, then  $f^{-1}$  is also a homomorphism.

It is clear from the definition of the partial order on Boolean algebras or lattices that every homomorphism  $f : K \rightarrow L$  also satisfies  $f(A) \leq f(B)$  for all  $A, B \in K$  such that  $A \leq B$ , i.e. it is in particular a poset homomorphism. In particular,  $f(0) = 0$  and  $f(1) = 1$ .

The converse of the above is generally not true, that is, it may happen that a poset homomorphism is not a lattice homomorphism.

**12.1.11. Fixed-point theorems.** Many practical problems lead to discussion on the existence and properties of fixed points of a mapping  $f : K \rightarrow K$  on a set  $K$ , i.e. of elements  $x \in K$  such that  $f(x) = x$ . The concepts of infima and suprema allows the derivation of very strong propositions of this type surprisingly easily. There follows here a classical theorem proved by Knaster and Tarski<sup>1</sup>:

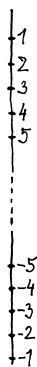


TARSKI'S FIXED-POINT THEOREM

**Theorem.** Let  $(K, \wedge, \vee)$  be a complete lattice and  $f : K \rightarrow K$  a poset homomorphism. Then,  $f$  has a fixed point, and the set of all fixed points of  $f$ , together with the restricted ordering from  $K$ , is again a complete lattice.

**PROOF.** Denote  $M = \{x \in K; x \leq f(x)\}$ . Since  $K$  has a least element,  $M$  is non-empty. Since  $f$  is order-preserving,  $f(M) \subseteq M$ . Moreover, denote  $z_1 = \sup M$ . Then, for  $x \in M, x \leq z_1$ , which means that  $f(x) \leq f(z_1)$ . At the same time,  $x \leq f(x)$ , hence  $f(z_1)$  is an upper bound for  $M$ . Then  $z_1 \leq f(z_1)$ , and also  $z_1 \in M$ , and hence  $f(z_1) \leq z_1$ . It follows that  $f(z_1) = z_1$ , so a fixed point is found.

<sup>1</sup>Knaster and Tarski proved this in the special case of the Boolean algebra of all subsets in a given set already in 1928, cf. Ann. Soc. Polon. Math. 6: 133–134. Much later in 1955, Tarski published the general result, cf. Pacific Journal of Mathematics. 5:2: 285–309. Alfred Tarski (1901-1983) was a renowned and influential Polish logician, mathematician and philosopher, who worked most of his active career in Berkeley, California. His elder colleague Bronisław Knaster (1893-1980) was also a Polish mathematician.



□

**12.A.40.** Give an example of an infinite chain which is a complete lattice.

**Solution.** We can take the set of real numbers together with  $-\infty, \infty$ , where  $-\infty$  is the least element (and thus the supremum of the empty set) and  $\infty$  is the greatest element (and thus the infimum of the empty set). The lattice suprema and infima are thus defined in accordance with these concepts in the real numbers. Moreover,  $-\infty$  is the infimum of subsets which are not bounded from below, and similarly  $\infty$  is the supremum of subsets which are not bounded from above. □

**12.A.41.** Decide whether the set of all convex subsets of  $\mathbb{R}^3$  is a lattice (with respect to suitably defined operations of suprema and infima). If so, is this lattice complete, distributive?

**Solution.** It is a lattice. The infimum is simply the intersection, since the intersection of convex subsets is again convex. The supremum is the convex hull of the union. It is clear that the lattice axioms are indeed satisfied for these operations (think this out!).

The lattice is complete, since the above operations work for infinite subsets as well, and clearly, the lattice has both a least element (the empty set) and a greatest element (the entire space).

However, the lattice is not distributive. For example, consider three unit balls  $B_1, B_2, B_3$  centered at  $[3, 0, 0], [-3, 0, 0], [0, 0, 0]$ , respectively. Then,

$$K_1 \vee (K_2 \wedge K_3) = K_1 \neq (K_1 \vee K_3) \wedge (K_1 \vee K_2).$$

□

**12.A.42.** Decide whether the set of all vector subspaces of  $\mathbb{R}^3$  is a lattice (with respect to suitably defined operations of suprema and infima). If so, is this lattice complete, distributive?

**Solution.** This is a lattice, infima correspond to intersections and suprema to sums of vector spaces (it is easy to verify that these operations satisfy the lattice axioms).

This lattice is complete (the operations work for infinite subsets as well, the least element is the zero-dimensional subspace, and the greatest element is the entire space).

However, it is not distributive (consider three lines in a plane). □



It is more difficult to verify the last statement of the theorem, namely that the set  $Z \subseteq K$  of all fixed points of  $f$  is a complete lattice. The greatest element  $z_1 = \max Z$  is found already. Analogously, using infimum and the property  $f(x) \leq x$  in the definition of  $M$ , we find the least element  $z_0 = \min Z$ .

Consider any non-empty set  $Q \subseteq Z$  and denote  $y = \sup Q$ . This supremum need not lie in  $Z$ . However, as seen shortly, the set has a supremum in  $Z$  with respect to the partial order in  $K$ , restricted to  $Z$ . For that purpose, denote  $R = \{x \in K; y \leq x\}$ . It is clear from the definitions that this set together with the partial order in  $K$ , restricted to  $R$ , is again a complete lattice, and that the restriction of  $f$  to  $R$  is again a poset homomorphism  $f|_R : R \rightarrow R$ . By the above,  $f|_R$  has a least fixed point  $\bar{y}$ . Of course,  $\bar{y} \in Z$ , and  $\bar{y}$  is the supremum of the fixed set  $Q$  with respect to the inherited order on  $Z$ . Note that it is possible that  $\bar{y} > y$ . Analogously, the infimum of any non-empty subset of  $Z$  can be found. Since the least and greatest elements are already found, the proof is finished. □

**Remark.** In the literature, one may find many variants of the fixed-point theorems, in various contexts. One of very useful variants is *Kleene's recursion theorem*, which can be determined from the theorem just proved and formulated as follows:



Consider a poset homomorphism  $f$  and a countable subset of  $K$  (using the notation of Tarski's fixed-point theorem), formed by the *Kleene chain*

$$0 \leq f(0) \leq f(f(0)) \leq \dots$$

Then, the supremum  $z$  of this subset cannot be greater than any fixed point of  $f$ . If  $y$  is a fixed point of  $f$ , it follows from  $0 \leq y$  that  $f(0) \leq f(y) = y$ , etc. Moreover, if  $f$  is assumed *continuous* in a certain sense of reasonably preserving suprema, then it can be shown that  $f(z)$  is also the supremum of this chain and hence is a fixed point. Therefore, it is the smallest fixed point. This theorem is called *Kleene's fixed-point theorem*. It has many applications in recursion theory, when discussing termination of algorithms, etc.

We omit details about the necessary "continuity" of mappings between posets and further generalizations.<sup>2</sup> We point out the added value to the general formulation of Tarski's theorem — the Kleene's theorem provides an iterative computational process approaching the fixed point with the given "seed", the minimal point.

<sup>2</sup>Stephen Cole Kleene (1909-1994) was a famous American mathematician working with Church, Turing, Post and others. The interested reader may consult full exposition of the above mentioned theorem in chapter 1 of the book: *Mathematical Theory of Domains*, Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, 1994, by V. Stoltenberg-Hansen, I. Lindström, E. R. Griffor.



**B. Rings**

**12.B.1.** Decide whether the set  $R$  with the operations  $\oplus, \odot$  form a ring, a commutative ring, an integral domain or a field:

- i)  $R = \mathbb{Z}, a \oplus b = a + b + 3, a \odot b = -3,$
- ii)  $R = \mathbb{Z}, a \oplus b = a + b - 3, a \odot b = a \cdot b - 1,$
- iii)  $R = \mathbb{Z}, a \oplus b = a + b - 1, a \odot b = a + b - a \cdot b,$
- iv)  $R = \mathbb{Q}, a \oplus b = a + b, a \odot b = b,$
- v)  $R = \mathbb{Q}, a \oplus b = a + b + 1, a \odot b = a + b + a \cdot b,$
- vi)  $R = \mathbb{Q}, a \oplus b = a + b - 1, a \odot b = a + b + a \cdot b.$

**Solution.**

- i) not a ring (but is a commutative rng),
- ii) not a ring,
- iii) an integral domain,
- iv) not a ring,
- v) a field, □
- vi) not a ring.

**12.B.2.** Prove that the subset  $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$  of the complex numbers is an integral domain. Is it a field?

**Solution.** Any subring of an integral domain must be an integral domain again. In this case, we are talking about a subset of the field  $\mathbb{C}$  (thus also an integral domain). Since the subset is closed with respect to all the operations (sum, additive inverse, multiplication) and contains both 0 and 1, it is indeed a subring. However, multiplicative inverses exist only for the numbers 1,  $i$ ,  $-1$ ,  $-i$  (these form the so-called subgroup of units – invertible elements), so it is not a field. □

**12.B.3.** In the ring of 2-by-2 matrices over the real numbers, consider the subring of matrices of the form  $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ .

Prove that this subring is isomorphic to  $\mathbb{C}$ .

**Solution.** We will show that the isomorphism is given by the mapping  $\varphi : \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \mapsto a + ib$ . The multiplication in the subring works as follows:

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \cdot \begin{pmatrix} c & -d \\ d & c \end{pmatrix} = \begin{pmatrix} ac - bd & -bc - ad \\ bc + ad & ac - bd \end{pmatrix},$$

and in  $\mathbb{C}$ , we have  $(a + ib)(c + id) = ac - bd + i(bc + ad)$ . Hence we can see that  $\varphi$  is a homomorphism with respect to multiplication. Since addition is defined componentwise,  $\varphi$  is a homomorphism to it as well. Moreover, this mapping is clearly both injective and surjective, thus it is an isomorphism. □

**12.B.4.** Prove that the identity is the only automorphism of the field of real numbers.

**Solution.** Consider an automorphism  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ . Clearly, it must satisfy  $\varphi(0) = 0$  and  $\varphi(1) = 1$ . Since  $\varphi$  respects addition, we must have for all positive integers  $n$  that  $\varphi(n) = \varphi(1 + 1 + \dots + 1) = n\varphi(1) = n$  and  $\varphi(-n) = -n$ . Since it respects multiplication, we must have for any integers

**12.1.12. Back to Boolean algebras.** When discussing propositional logic, there is the problem of what exactly are the elements of the corresponding Boolean algebra. Formally, they are defined as the classes of equivalent propositions. In other words, we work with truth-value functions for propositions with a given number of arguments. There is the problem of recognizing propositions which are equivalent in this sense. There is the question of whether every function  $(\mathbb{Z}_2)^n \rightarrow \mathbb{Z}_2$  can be defined in terms of the basic logical operations. Clearly all such functions form a Boolean algebra, since their values are in the Boolean algebra  $\mathbb{Z}_2$ .



Similarly, there is the problem of deciding whether or not two systems of switches can have the same function. Just as for propositions, a system consisting of  $n$  switches corresponds to a function  $(\mathbb{Z}_2)^n \rightarrow \mathbb{Z}_2$ . There are  $2^{2^n}$  such functions. A Boolean algebra can be naturally defined on these functions (again using the fact that the function values are in the Boolean algebra  $\mathbb{Z}_2$ ).

We summarize a few such questions:

SOME BASIC QUESTIONS

*Question 1:* Are all finite Boolean algebras  $(K, \wedge, \vee)$  defined on sets  $K$  with  $2^n$  elements?

*Question 2:* Can each function  $(\mathbb{Z}_2)^n \rightarrow \mathbb{Z}_2$  be the truth function of some logical expression built of  $n$  elementary propositions and the logical operators?

*Question 3:* How to recognize whether two such expressions represent the same function?

*Question 4:* Can each function  $(\mathbb{Z}_2)^n \rightarrow \mathbb{Z}_2$  be realized by some switch board with  $n$  switches?

*Question 5:* How to recognize whether two switchboards represent the same function?

All these questions are answered by finding the normal form of every element of a general Boolean algebra. This is achieved by writing it as the join of certain particularly simple elements. By comparing the normal forms of any pair of elements, it is easily determined whether or not they are the same.

This helps to classify all finite Boolean algebras, giving the affirmative answer to question 1.

**12.1.13. Atoms and normal forms.** First, define the “simplest” elements of a Boolean algebra:

ATOMS IN A BOOLEAN ALGEBRA

Let  $K$  be a Boolean algebra. An element  $A \in K, A \neq 0$ , is called an *atom* if and only if for all  $B \in K, A \wedge B = A$  or  $A \wedge B = 0$ .

In other words,  $A \neq 0$  is an atom if and only if there are only two elements  $B$  such that  $B \leq A$ , namely  $B = 0$  and  $B = A$ .

Note that 0 is not considered an atom, just as the integer 1 is not considered a prime. Let us remark that infinite Boolean algebras may contain no atoms at all.

$p, q$  ( $q \neq 0$ ) that  $\varphi(p) = \varphi(q \cdot \frac{p}{q}) = \varphi(p) \cdot \varphi(\frac{p}{q})$ . Hence,  $\varphi(\frac{p}{q}) = \frac{p}{q}$ , i. e.,  $\varphi(r) = r$  for all rational numbers  $r$ .

Consider a positive number  $x \in \mathbb{R}$ . Then,  $\varphi(x) = \varphi(\sqrt{x^2}) = \varphi(\sqrt{x})^2 \geq 0$ . Thus, for any  $x, y \in \mathbb{R}$  such that  $x < y$ , we must have  $\varphi(x) < \varphi(y)$ . Now, assume that  $\varphi$  is not the identity, i. e., there exists a  $z \in \mathbb{R}$  such that  $\varphi(z) \neq z$ . We can assume without loss of generality that  $\varphi(z) < z$ . Since  $\mathbb{Q}$  is dense in  $\mathbb{R}$ , there exists an  $r \in \mathbb{Q}$  for which  $\varphi(z) < r < z$ . However, we know that  $\varphi(r) = r$ , which means that  $r < z$  implies  $\varphi(r) < \varphi(z)$ . Altogether, we get the wanted contradiction  $\varphi(z) < \varphi(r) < \varphi(z)$ .  $\square$

**12.B.5.** Let  $p$  be a prime and  $R$  a ring which contains  $p^2$  elements. Prove that  $R$  is commutative.

**Solution.** Since  $(R, +)$  is a finite commutative group with  $p^2$  elements, it is by 12.4.8 isomorphic to either  $\mathbb{Z}_{p^2}$  or  $\mathbb{Z}_p \times \mathbb{Z}_p$ . In the first case,  $(R, +)$  is cyclic, so there exists an element  $x \in R$  such that each element of  $R$  is of the form  $nx$  for some  $1 \leq n \leq p^2$ . Since all these elements commute, we get that the entire  $R$  is commutative.

In the second case, each element (except 0) must have order  $p$  with respect to addition. Let  $x \in R$  be any element that is not in the additive subgroup generated by 1. Then, each element of  $R$  is of the form  $m + nx$ , where  $1 \leq m, n \leq p$ . Again, all these elements commute, so  $R$  is commutative.  $\square$

**12.B.6.** Find the inverses of 17, 18, and 19 in  $(\mathbb{Z}_{131}^*, \cdot)$  (the group of all invertible elements in  $\mathbb{Z}_{131}$  with multiplication).

**Solution.** Applying the Euclidean algorithm, we get

$$\begin{aligned} 131 &= 7 \cdot 17 + 12, \\ 17 &= 1 \cdot 12 + 5, \\ 12 &= 2 \cdot 5 + 2, \\ 5 &= 2 \cdot 2 + 1. \end{aligned}$$

Therefore,  $1 = 5 - 2 \cdot 2 = 5 - 2(12 - 2 \cdot 5) = 5 \cdot 5 - 2 \cdot 12 = 5 \cdot (17 - 12) - 2 \cdot 12 = 5 \cdot 17 - 7 \cdot 12 = 5 \cdot 17 - 7 \cdot (131 - 7 \cdot 17) = 54 \cdot 17 - 7 \cdot 131$ . The inverse of 17 is 54. Similarly,  $[18]^{-1} = 51$  and  $[19]^{-1} = 69$ .  $\square$

**12.B.7.** Find the inverse of  $[49]_{\mathbb{Z}_{253}}$  in  $\mathbb{Z}_{253}$  ○

**12.B.8.** Find the inverse of  $[37]_{\mathbb{Z}_{208}}$  in  $\mathbb{Z}_{208}$ . ○

**12.B.9.** Find the inverse of  $[57]_{\mathbb{Z}_{359}}$  in  $\mathbb{Z}_{359}$ . ○

**12.B.10.** Find the inverse of  $[17]_{\mathbb{Z}_{40}}$  in  $\mathbb{Z}_{40}$ . ○

### C. Polynomial rings

**12.C.1. Eisenstein's irreducibility criterion** This criterion provides a sufficient condition for a polynomial over  $\mathbb{Z}$  to be irreducible over  $\mathbb{Q}$  (which is the same as to be irreducible over  $\mathbb{Z}$ ):

Let

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

be a polynomial over  $\mathbb{Z}$  and  $p$  be a prime such that

- $p$  divides  $a_j, j = 0, \dots, n - 1,$

The situation is very simple in the Boolean algebra of all subsets of a given finite set  $M$ . Clearly, the atoms are precisely the singletons  $A = \{x\}$ . For every subset  $B$ , either  $A \wedge B = A$  (if  $x \in B$ ) or  $A \wedge B = 0$  (if  $x \notin B$ ). The requirements fail whenever there is more than one element in  $A$ .

Next, consider which elements are atoms in the Boolean algebra of functions of the switch boards with  $n$  switches  $A_1, \dots, A_n$ . It can be easily verified that there are  $2^n$  atoms, which are of the form  $A_1^{\sigma_1} \wedge \dots \wedge A_n^{\sigma_n}$ , where either  $A_i^{\sigma_i} = A_i$  or  $A_i^{\sigma_i} = A_i'$ .

The infimum  $\varphi \wedge \psi$  of functions  $\varphi$  and  $\psi$  is the function whose values are given by the products of the corresponding values in  $\mathbb{Z}_2$ . Therefore,  $\varphi \leq \psi$  if  $\varphi$  takes the value 1 in  $\mathbb{Z}_2$  only on arguments where  $\psi$  also has value 1. Hence in the Boolean algebra of truth-value functions, a function  $\varphi$  is an atom if and only if  $\varphi$  returns 1 in  $\mathbb{Z}_2$  for exactly one of the  $2^n$  possible choices of arguments. All these functions can be created in the above mentioned manner.

Now, the promised theorem can be formulated. While this one is called the *disjunctive normal form*, there is also the opposite version with the suprema and infima interchanged (the *conjunctive normal form*).

#### DISJUNCTIVE NORMAL FORM

**Theorem.** Each element  $B$  of a finite Boolean algebra  $(K, \wedge, \vee)$  can be written as a supremum of atoms

$$B = A_1 \vee \dots \vee A_k.$$

*This expression is unique up to the order of the atoms.*

The proof takes several paragraphs, but the basic idea is quite simple: Consider all atoms  $A_1, A_2, \dots, A_k$  in  $K$  which are less or equal to  $B$ . From the properties of the order on  $K$ , (see 12.1.8(3)) it follows that



$$Y = A_1 \vee \dots \vee A_k \leq B.$$

The main step of the proof is to verify that  $B \wedge Y' = 0$ , which by 12.1.8(4) guarantees that  $B \leq Y$ . That proves the equality  $B = Y$ .

**12.1.14. Three useful claims.** We derive several technical properties of atoms, in order to complete the proof of the theorem on disjunctive normal form. We retain the notation of the previous subsection.

**Proposition.** (1) If  $Y, X_1, \dots, X_\ell$  are atoms in  $K$ , then  $Y \leq X_1 \vee \dots \vee X_\ell$  if and only if  $Y = X_i$  for some  $i = 1, \dots, \ell$ .

(2) For each  $Y \in K, Y \neq 0$ , there is an atom  $X \in K$  such that  $X \leq Y$ .

(3) If  $X_1, \dots, X_r$  are precisely all the atoms of  $K$ , then  $Y = 0$  if and only if  $Y \wedge X_i = 0$  for all  $i = 1, \dots, r$ .

**PROOF.** (1) If the inequality of the proposition holds, then

$$Y \wedge (X_1 \vee \dots \vee X_\ell) = Y.$$

- $p$  does not divide  $a_n$ ,
- $p^2$  does not divide  $a_0$ .

Then,  $f(x)$  is irreducible over  $\mathbb{Z}$  ( $\mathbb{Q}$ ). Prove this criterion.

○

**12.C.2.** Factorize over  $\mathbb{C}$  and  $\mathbb{R}$  the polynomial

$$x^4 + 2x^3 + 3x^2 + 2x + 1.$$

**Solution.** This polynomial can be factorized either by looking for multiple roots or as a reciprocal equation:

- Let us compute the greatest common divisor of the polynomial and its derivative  $4x^3 + 6x^2 + 6x + 2$ , using the Euclidean algorithm. The greatest common divisor is given in any ring up to a multiple by a unit, and during the Euclidean algorithm, we may multiply the partial results by units of the ring. In the case of a polynomial ring over a field of scalars, the units are exactly the non-zero scalars. We perform the multiplication in the way to avoid calculations with fractions as much as possible.

$$2x^4 + 4x^3 + 6x^2 + 4x + 2 : 2x^3 + 3x^2 + 3x + 1 = x + \frac{1}{2}$$

$$\begin{aligned} 2x^4 + 3x^3 + 3x^2 + x \\ x^3 + 3x^2 + 3x + 2 \\ x^3 + \frac{3}{2}x^2 + \frac{3}{2}x + \frac{1}{2} \\ \frac{3}{2}x^2 + \frac{3}{2}x + \frac{3}{2} \end{aligned}$$

Further, we divide the polynomial  $2x^3 + 3x^2 + 3x + 1$  by the remainder  $\frac{3}{2}x^2 + \frac{3}{2}x + \frac{3}{2}$  (multiplied by the unit  $\frac{2}{3}$ )

$$\begin{aligned} 2x^3 + 3x^2 + 3x + 1 : x^2 + x + 1 = 2x + 1 \\ 2x^3 + 2x^2 + 2x \\ x^2 + x + 1 \end{aligned}$$

The roots of the greatest common divisor of the original polynomial and its derivative are exactly the multiple roots of the original polynomial. In this case, the roots of  $x^2 + x + 1$  are  $-\frac{1}{2} \pm i\sqrt{3}/2$ , which are thus double roots of the original polynomial. The factorization over  $\mathbb{C}$  is thus to root factors (this is always the case over  $\mathbb{C}$ , as stated by the fundamental theorem of algebra):

$$\begin{aligned} x^4 + 2x^3 + 3x^2 + 2x + 1 = \\ = \left(x + \frac{1}{2} - i\frac{\sqrt{3}}{2}\right)^2 \cdot \left(x + \frac{1}{2} + i\frac{\sqrt{3}}{2}\right)^2. \end{aligned}$$

The factorization over  $\mathbb{R}$  can be obtained by multiplying the factors corresponding to pairs of complex-conjugated roots of the polynomial (verify that such a product must always result in a polynomial with real coefficients!):

$$x^4 + 2x^3 + 3x^2 + 2x + 1 = (x^2 + x + 1)^2.$$

- Let us solve the equation

$$x^4 + 2x^3 + 3x^2 + 2x + 1 = 0.$$

By distributivity, the equality can be rewritten as

$$(Y \wedge X_1) \vee \dots \vee (Y \wedge X_\ell) = Y.$$

However, for all  $i$  either  $Y \wedge X_i = 0$  or  $Y \wedge X_i = X_i$ . If all these intersections are 0, then  $Y = 0$ . Thus, there is an  $i$  for which  $Y \wedge X_i = X_i$ . Since  $Y$  is also an atom, the desired equality  $Y = X_i$  is proved.

The other implication is trivial.

(2) If  $Y$  is an atom itself, choose  $X = Y$ . If  $Y$  is not an atom, then it follows from the definition that there must exist a non-zero element  $Z_1 \neq Y$  for which  $Z_1 \leq Y$ . If  $Z_1$  is not an atom either, then similarly find a  $Z_2 \leq Z_1$ , etc., leading to a sequence of pairwise distinct elements

$$\dots \leq Z_k \leq Z_{k-1} \leq \dots \leq Z_1 \leq Y,$$

which cannot be infinite since the entire Boolean algebra  $K$  is finite. Therefore, it must end with an atom  $Z_k$ .

(3) Assume that  $Y \wedge X_i = 0$  for all indices  $i$ . If  $Y \neq 0$ , then due to the above claim, there must exist an atom  $X_j$  for which  $X_j \wedge Y = X_j$ , which is a contradiction.

The other implication is trivial. □

**12.1.15. Completion of the proof of theorem 12.1.13.**

Write

$$Y = A_1 \vee \dots \vee A_k \leq B,$$

where  $A_i$  are all the atoms in  $K$  which are less than or equal to  $B$ . Compute

$$B \wedge Y' = B \wedge (A_1 \vee \dots \vee A_k)' = B \wedge A_1' \wedge \dots \wedge A_k'.$$

If an atom  $A = A_i$  is contained in the join  $Y$ , then  $B \wedge Y' \wedge A = 0$ . However, if  $A$  is an atom which does not occur in  $Y$ , then also  $B \wedge Y' \wedge A = 0$ , since  $Y$  contains exactly those atoms which are  $\leq B$ . Hence  $B \wedge Y' \wedge A = 0$  for all atoms  $A$  in  $K$ .

Thus it is proved that the intersection of  $B \wedge Y'$  and any atom is zero, which means that it must be zero itself, by the third claim in the latter proposition. Therefore,  $B \leq Y$  (cf. 12.1.8(4)). The definition of  $Y$  implies  $Y \leq B$ , so the anti-symmetry of the order implies that  $B = Y$ .

It remains to prove the uniqueness of the expression, up to order. Thus, suppose  $B$  can be written in two ways as

$$B = A_1 \vee \dots \vee A_k = \tilde{A}_1 \vee \dots \vee \tilde{A}_\ell.$$

Since each  $A_i$  satisfies  $A_i \leq B$ , the first claim in the proposition above ensures it must equal one of the  $\tilde{A}_j$ . Repeating this argument gives the desired uniqueness and finishes the proof.

**12.1.16. Classification.** To end the discussion of Boolean algebras, we prove that all the examples of finite Boolean algebras (of given size) are isomorphic.



In particular, each of the  $2^{2^n}$  truth-value functions for  $n$  atomic propositions can be expressed as an appropriate proposition, just like each of the  $2^{2^n}$  switch board functions can be defined in terms of  $n$  suitably arranged switches. In both cases, the algebra in question behaves the

Dividing by  $x^2$  and substituting  $t = x + \frac{1}{x}$ , we get the equation

$$t^2 + 2t + 1 = 0$$

with double root  $-1$ . Now, substituting this into the definition of  $t$ , we get the known equation  $x^2 + x + 1 = 0$ , which was solved above.

**Remark.** Let us remark that the only irreducible polynomials over  $\mathbb{R}$  are linear polynomials and quadratic polynomials with negative discriminant. This also follows from the reasonings in the above exercise.

**12.C.3.** Factorize the polynomial  $x^5 + 3x^3 + 3$  to irreducible factors over

- i)  $\mathbb{Q}$ ,
- ii)  $\mathbb{Z}_7$ .

**Solution.**

- i) By Eisenstein's criterion, the given polynomial is irreducible over  $\mathbb{Z}$  and  $\mathbb{Q}$  (we use the prime 3).
- ii)  $(x - 1)^2(x^3 + 2x^2 - x + 3)$ . Using Horner's scheme, for instance, we find the double root 1. When divided by the polynomial  $(x - 1)^2$ , we get  $(x^3 + 2x^2 - x + 3)$ , which has no roots over  $\mathbb{Z}_7$ . Since it is only of degree 3, this means that it must be irreducible (if it were reducible, one of the factors would have to be linear, which means that the cubic polynomial  $(x^3 + 2x^2 - x + 3)$  would have a root).

**12.C.4.** Factorize the polynomial  $x^4 + 1$  over

- $\mathbb{Z}_3$ ,
- $\mathbb{C}$ ,
- $\mathbb{R}$ .

**Solution.**

- $(x^2 + x + 2)(x^2 + 2x + 2)$
- The roots are the fourth roots of  $-1$ , which lie in the complex plane on the unit circle, and their arguments are  $\pi/4$ ,  $\pi/4 + \pi/2$ ,  $\pi/4 + \pi$ , and  $\pi/4 + 3\pi/2$  i. e., they are the numbers  $\pm\sqrt{2}/2 \pm i\sqrt{2}/2$ . Thus, the factorization is

$$\left(x - \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}\right) \left(x - \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}\right) \left(x + \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}\right) \left(x + \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}\right)$$

- Multiplying the root factors of complex-conjugated roots in the factorization over  $\mathbb{C}$ , we get the factorization over  $\mathbb{R}$ : □

$$(x^2 - \sqrt{2}x + 1)(x^2 + \sqrt{2}x + 1) ..$$

**12.C.5.** Find a polynomial with rational coefficients of the lowest degree possible which has  $\sqrt[2007]{2}$  as a root.

**Solution.**  $P(x) = x^{2007} - 2$ . Let us show that there is no polynomial of lower degree with root  $\sqrt[2007]{2}$ : Let  $Q(x)$  be a non-zero polynomial of the lowest degree with root  $\sqrt[2007]{2}$ . Then,  $\deg Q(x) \leq 2007$ . Let us divide  $P(x)$  by  $Q(x)$  with remainder:  $P(x) = Q(x) \cdot D(x) + R(x)$ , where  $D(x)$  is the quotient and  $R(x)$  is the remainder, and either  $\deg R(x) < \deg Q(x)$  or  $R(x) = 0$ . Substituting the number  $\sqrt[2007]{2}$  into the last equation, we can see that  $\sqrt[2007]{2}$  is also a root of  $R(x)$ . By the definition of  $Q(x)$ , this means that  $R(x)$  must be the zero polynomial, which means that  $Q(x)$  divides  $P(x)$ . However,  $P(x)$  is irreducible (by Eisenstein's criterion for 2), so its only

same way as the Boolean algebra of all subsets of a given  $2^n$ -element set.

Moreover, each of these expressions can be written in a unique normal form, so it can be decided algorithmically whether two switch boards have the same behaviour without comparing their values for all  $2^n$  possible inputs (which on the other hand might still be faster, in particular the resulting normal formula tends to be exponentially large).

**Theorem.** Every finite Boolean algebra is isomorphic to the Boolean algebra  $K = 2^M$  where  $M$  is the set of atoms in  $K$ .

**PROOF.** The idea of the proof is quite straightforward. Every isomorphism of a Boolean algebra  $(K, \wedge, \vee)$  must map atoms to atoms. Let  $M$  be the set of all atoms in  $K$  and consider the Boolean algebra  $(2^M, \cap, \cup)$ . This defines a natural correspondence between the atoms of  $K$  and the atoms of  $2^M$ .

Next, use the disjunctive normal form to extend the mapping to all of  $K$ . Each element  $X \in K$  can be written uniquely (up to order) as a join of atoms:

$$X = A_1 \vee \dots \vee A_k$$

Define the function  $f : K \rightarrow 2^M$  by

$$f(X) = f(A_1) \cup \dots \cup f(A_k) = \{A_1, \dots, A_k\},$$

as the union of the singletons  $A_i \subseteq M$  that occur in the expression.

The uniqueness of the normal form implies that  $f$  is a bijection. It remains to show that it is a homomorphism of the Boolean algebras.

Let  $X, Y \in K$ . The normal form of their supremum contains exactly the atoms which occur in at least one of  $X, Y$ ; while the infimum involves just those atoms which occur in both. This verifies that  $f$  preserves the operations  $\wedge$  and  $\vee$ . As for the complements, note that an atom  $A$  occurs in the normal form of  $X'$  if and only if  $X \wedge A = 0$ . Hence  $f$  preserves complements, which finishes the proof. □

The classification of infinite Boolean algebras is far more complicated. It is not the case that each would be isomorphic to the Boolean algebra of all subsets of an appropriate set  $M$ . However, every Boolean algebra is isomorphic to a Boolean subalgebra of a Boolean algebra  $2^M$  for an appropriate set  $M$ . This result is known as *Stone's representation theorem*<sup>3</sup>.

## 2. Elements of Logic

This section provides a quick introduction to logic. Mastering the calculus of Boolean algebras will become useful, but we have to start from scratch.

<sup>3</sup>The American mathematician Marshall Harvey Stone (1903 – 1989) proved this theorem in 1936 when dealing with the spectral theory of operators on Hilbert spaces, required for analysis and topology. Nowadays, it belongs to standard material in advanced textbooks.

non-trivial divisor is itself (up to multiplication by a unit of the polynomial ring over  $\mathbb{Q}$ , i. e. a non-zero rational constant). Thus, we have  $Q(x) = P(x)$  up to multiplication by a unit. For instance, the polynomial  $\frac{1}{3}x^{2007} - \frac{2}{3}$  also satisfies the stated conditions. However, if we require the polynomial to be monic (i. e., with leading coefficient 1), then the only solution is the mentioned polynomial  $P(x)$ .  $\square$

**12.C.6.** Find all irreducible polynomials of degree at most 2 over  $\mathbb{Z}_3$ .

**Solution.** By definition, all linear polynomials are irreducible. As for quadratic irreducible polynomials, the easiest way is to simply enumerate them all and leave out the reducible ones, i. e. those which are a product of two linear polynomials. The reducible polynomials are  $(x + 1)^2 = x^2 + 2x + 1$ ,  $(x + 2)^2 = x^2 + x + 1$ ,  $(x + 1)(x + 2) = x^2 + 2$ ,  $x^2$ ,  $x(x + 1) = x^2 + x$ ,  $x(x + 2) = x^2 + 2x$ . (It suffices to consider monic polynomials, since the non-monic can be obtained by multiplication by 2.) The remaining quadratic polynomials over  $\mathbb{Z}_3$  are irreducible; these are  $x^2 + 2x + 2$ ,  $x^2 + x + 2$ ,  $x^2 + 1$ .  $\square$

**12.C.7.** Decide whether the following polynomial is irreducible over  $\mathbb{Z}_3$ ; if not, factorize it:

$$x^4 + x^3 + x + 2.$$

**Solution.** Evaluating the polynomial at 0, 1, 2, we find that it has no root in  $\mathbb{Z}_3$ . This means that it is either irreducible or a product of two quadratic polynomials. Assume it is reducible. Then, we may assume without loss of generality that it is a product of two monic polynomials (the only other option is that it is a product of two polynomials with leading coefficients equal to 2 – then both can be multiplied by 2 in order to become monic). Thus, let us look for constants  $a, b, c, d \in \mathbb{Z}_3$  so that

$$\begin{aligned} x^4 + x^3 + x + 2 &= (x^2 + ax + b)(x^2 + cx + d) = \\ &= x^4 + (a + c)x^3 + (ac + b + d)x^2 + \\ &\quad + (ad + bc)x + bd. \end{aligned}$$

Comparing the coefficients of individual power of  $x$ , we get the following system of four equations in four variables:

$$\begin{aligned} 1 &= a + c, \\ 0 &= ac + b + d, \\ 1 &= ad + bc, \\ 2 &= bd. \end{aligned}$$

From the last equation, we get that one of the numbers  $b, d$  is equal to 1 and the other one to 2. Thanks to symmetry of the system in the pairs  $(a, b)$  and  $(c, d)$ , we can choose  $b = 1$ ,  $d = 2$ . From the second equation, we get  $ac = 0$ , i. e., at least one of the numbers  $a, c$  is 0. From the first equation, we get that the other one is 1. From the third equation, we get  $2a + c = 1$ , i. e.,  $a = 0, c = 1$ . Altogether,

$$x^4 + x^3 + x + 2 = (x^2 + 1)(x^2 + x + 2). \quad \square$$

**12.2.1. From set algebra to logical propositions.** Let us recall that in 12.1.2, we abstracted Boolean algebras from the properties of set operations; in particular, the  $\vee$  operation corresponded to the union of sets. We understand sets as collections of elements with some property. So, if an element  $x$  is in the union of sets  $C$  and  $B$ , then it has "property  $C$ " or "property  $B$ ." Furthermore, if we know that element  $x$  does not have "property  $C$ ," then we know for sure that it has "property  $B$ ." In other words, from the fact that element  $x$  has "property  $C$ " or "property  $B$ ," but not "property  $C$ ," we can conclude that element  $x$  definitely has "property  $B$ ." Symbolically,

$$\frac{x \in C \cup B}{x \notin C} \quad , \quad \frac{}{x \in B} \quad ,$$

or more succinctly using Boolean operations,

$$\frac{C \vee B}{C}$$

This simple observation can be reformulated as follows: "From  $C \vee B$  and  $C'$ , it logically follows that  $B$ ." For the sake of simplicity, let's denote  $A = C'$ , which means  $C = A'$ , and rewrite the previous observation as:

$$(1) \quad \frac{A' \vee B}{A}$$

and read it as: "From  $A' \vee B$  and  $A$ , it necessarily implies  $B$ ."

Next, let's perform three simple Boolean calculations:

$$\begin{aligned} A' \vee (B' \vee A) &= (A' \vee A) \vee B = 1 \vee B = 1, \\ (A' \vee B)' \vee ((A' \vee B)') \vee A &= (A' \vee B)' \vee ((A \wedge B) \vee A') \\ &= (A' \vee B)' \vee (B \vee A') = 1, \\ (A' \vee B)' \vee (A' \vee C) &= (A \wedge B') \vee A' \vee C \\ &= (A \vee A' \vee C) \wedge (B' \vee A' \vee C) = 1 \wedge (A' \vee B' \vee C) \\ &= (A' \vee (B' \vee C)). \end{aligned}$$

Now, we will use the axioms (8) of Boolean algebras, and we obtain the following formulas:

$$(2) \quad \begin{aligned} A' \vee (B' \vee A) &= 1, \quad A' \vee A = 1 \\ (A' \vee B)' \vee ((A' \vee B)') \vee A &= 1, \\ (A' \vee (B' \vee C))' \vee ((A' \vee B)' \vee (A' \vee C)) &= 1. \end{aligned}$$

We already noticed the connections between Boolean algebras and propositional logic in paragraph 12.1.5. In this sense, we can see that the left sides of the previous equations express true propositions. Now, let's transition more thoroughly from the "Boolean world" to the "world of logic." We will do this by replacing uppercase letters with lowercase letters, replacing the symbol  $\vee$  with the symbol  $\rightarrow$ , and changing the "postfix" symbol  $'$  to the "prefix" symbol  $\neg$ . So, instead of  $A' \vee B$ , we will write  $a \rightarrow b$ , and instead of  $A'$ , we will write  $\neg a$ ."

**12.C.8.** For any odd prime  $p$ , find all roots of the polynomial

$$P(x) = x^{p-2} + x^{p-3} + \cdots + x + 2$$

over the field  $\mathbb{Z}_p$ .

**Solution.** Considering the equality

$$x^{p-1} - 1 = (x - 1)(P(x) - 1),$$

we can see that all numbers of  $\mathbb{Z}_p$ , except 0 and 1, are roots of  $P(x) - 1$ , so they cannot be roots of  $P(x) + 1$ . Clearly, 0 is never a root of  $P(x)$ , and 1 is always a root, which means that it is the only root.  $\square$

**12.C.9.** Factorize the polynomial  $p(x) = x^2 + x + 1$  in  $\mathbb{Z}_5[x]$  and  $\mathbb{Z}_7[x]$ .

**Solution.** Irreducible in  $\mathbb{Z}_5[x]$ ;  $p(x) = (x - 2)(x - 4)$  in  $\mathbb{Z}_7[x]$ .  $\square$

**12.C.10.** Factorize the polynomial  $p(x) = x^6 - x^4 - 5x^2 - 3$  in  $\mathbb{C}[x]$ ,  $\mathbb{R}[x]$ ,  $\mathbb{Q}[x]$ ,  $\mathbb{Z}[x]$ ,  $\mathbb{Z}_5[x]$ ,  $\mathbb{Z}_7[x]$ , knowing that it has a multiple root.

**Solution.** Applying the Euclidean algorithm, we find out that the greatest common divisor of  $p$  and its derivative  $p'$  is  $x^2 + 1$ . Dividing the polynomial  $p(x)$  twice by this factor, we get

$$p(x) = (x^2 + 1)^2(x^2 - 3).$$

Clearly, these factors are irreducible in the rings  $\mathbb{Q}[x]$  and  $\mathbb{Z}[x]$ .

In  $\mathbb{C}[x]$ , we can always factorize a polynomial to linear factors. In this case, it suffices to factorize  $x^2 + 1$ , which is easy:  $x^2 + 1 = (x + i)(x - i)$ . The factor  $x^2 - 3$  is equal to  $(x - \sqrt{3})(x + \sqrt{3})$  even in  $\mathbb{R}[x]$ . Thus, in  $\mathbb{C}[x]$ , we have

$$p(x) = (x + i)^2(x - i)^2(x - \sqrt{3})(x + \sqrt{3}),$$

while in  $\mathbb{R}[x]$ , we have

$$p(x) = (x^2 + 1)^2(x - \sqrt{3})(x + \sqrt{3}).$$

In  $\mathbb{Z}_5[x]$ , the polynomial  $x^2 + 1$  has roots  $\pm 2$ , and the polynomial  $x^2 - 3$  has no roots, which means that

$$p(x) = (x - 2)^2(x + 2)^2(x^2 - 3).$$

In  $\mathbb{Z}_7[x]$ , neither polynomial has a root, so the factorization to irreducible factors is identical to that in  $\mathbb{Q}[x]$  and  $\mathbb{Z}[x]$ .  $\square$

**12.C.11.** Knowing that the polynomial  $p = x^6 + x^5 + 4x^4 + 2x^3 + 5x^2 + x + 2$  has multiple root  $x = i$ , factorize it to irreducible polynomials over  $\mathbb{C}[x]$ ,  $\mathbb{R}[x]$ ,  $\mathbb{Z}_2[x]$ ,  $\mathbb{Z}_5[x]$ , and  $\mathbb{Z}_7[x]$ . Divide the polynomial  $q = x^2y^2 + y^2 + xy + x^2y + 2y + 1$  by the irreducible factors of  $p$  in  $\mathbb{R}[x]$ , and use the result to solve the system of polynomial equations  $p = q = 0$  over  $\mathbb{C}$ .

**Solution.**  $p = (x^2 + 1)^2(x^2 + x + 2)$ , in  $\mathbb{Z}_2$ :  $p = x(x + 1)^5$ , in  $\mathbb{Z}_5$ :  $p = (x - 2)^2(x + 2)^2(x^2 + x + 2)$ , in  $\mathbb{Z}_7$ :  $p = (x^2 + 1)^2(x + 4)^2$ . For the second polynomial, we get  $q = (y^2 + y)(x^2 + x + 2) - y^2(x + 1) + 1$  and  $q = (y^2 + y)(x^2 + 1) + y(x + 1) + 1$ . Thus, if  $x = \alpha$  is a root of  $x^2 + x + 2$ , i. e.,  $\alpha = -\frac{1}{2} \pm \frac{1}{2}i\sqrt{7}$ ,

**12.2.2. Classical Propositional Logic.** The alphabet of classical propositional logic consists of lowercase Latin letters  $a, b, c, \dots$  possibly with subscripts  $a_1, a_2, a_3, \dots$ , special symbols  $\neg, \rightarrow$ , and parentheses  $(, )$ .

These letters are called "atomic propositions" or "propositional variables." We can think of them as declarative statements that can be either true or false. (Truth in this context is understood in the usual vague sense; the concept of truth has not been precisely defined yet.) The symbols  $\neg, \rightarrow$  are called "operators" or "connectives," sometimes also "propositional functors."

#### WHAT ARE THE PROPOSITIONS?

**Definition.** "Propositions" are defined by the following rules:

- (i) Every atomic proposition is a proposition.
- (ii) If  $A$  is a proposition, then  $\neg A$  is also a proposition.
- (iii) If  $A$  and  $B$  are propositions, then  $(A \rightarrow B)$  is also a proposition.

In this definition, uppercase letters are actually parameters that represent any proposition.

**Example.**  $a, b, \neg a, (a \rightarrow b)$  are propositions;  $A, B, C$  are parameters that represent arbitrary propositions.

A statement created according to rule (ii) will be called *negation of statement A*. We can read it as 'not A,' 'non-A,' 'A is not true,' and similar expressions. A statement created according to rule (iii) is called an *implication*, where statement  $A$  is called the *antecedent of the implication*, and statement  $B$  is called the *consequent of the implication*. We can read it as 'A logically implies B,' 'if A, then B,' 'A is a sufficient condition for B,' 'B is a necessary condition for A,' 'A implies B,' and similar expressions.

If there is no risk of misunderstanding, we will omit the outer parentheses in statements created according to rule (iii). The last statement from the previous example will then have the form  $\neg((b \rightarrow a) \rightarrow \neg a) \rightarrow (b \rightarrow a)$ .

We consider the alphabet of propositional calculus as primitive concepts. From them, we construct, delineate, and define further concepts using rules. In Kantian terminology, atomic statements are synthetic, while other statements are analytical—they can be broken down into atomic statements.

Statements are understood as declarative sentences that are clear enough to decide whether they are true or not. Since 'truth' is still a very vague concept, we will use the word 'validity.' However, we do not yet know any valid statements. We will now declare as valid statements those that are prescribed by the Boolean expressions on the left sides of equations (2).

then  $y = \frac{1}{\sqrt{1+\alpha}}$ . If  $x = \beta$  is a root of  $x^2 + 1$ , i. e.,  $\beta = \pm i$ , then  $y = -\frac{1}{1+\beta}$  □

**12.C.12.** Factorize the following polynomial to irreducible polynomials in  $\mathbb{R}[x]$  and in  $\mathbb{C}[x]$ :

$$4x^5 - 8x^4 + 9x^3 - 7x^2 + 3x - 1.$$

**12.C.13.** Factorize the following polynomial to irreducible polynomials in  $\mathbb{R}[x]$  and in  $\mathbb{C}[x]$ :

$$x^5 + 3x^4 + 7x^3 + 9x^2 + 8x + 4.$$

**12.C.14.** Factorize  $x^4 - 4x^3 + 10x^2 - 12x + 9$  to irreducible polynomials in  $\mathbb{R}[x]$  and in  $\mathbb{C}[x]$ :

**12.C.15.** Decide whether the following polynomial over  $\mathbb{Z}_3$  is irreducible; if not, factorize it to irreducible polynomials:

$$x^5 + x^2 + 2x + 1.$$

**12.C.16.** Decide whether the following polynomial over  $\mathbb{Z}_3$  is irreducible; if not, factorize it to irreducible polynomials:

$$x^4 + 2x^3 + 2.$$

**12.C.17.** Find all monic quadratic irreducible polynomials over  $\mathbb{Z}_5$ .

**Solution.** We write out all monic quadratic polynomials over  $\mathbb{Z}_5$  and exclude those which are not irreducible, i. e., have a root:

$$x^2 \pm 2, x^2 \pm x + 2, x^2 \pm 2x - 2, x^2 - x \pm 1, x^2 \pm 2x - 1.$$

**D. Rings of multivariate polynomials**

**12.D.1.** Find the remainder of the polynomial  $x^3y + x + yz + yz^4$  with respect to the basis  $(x^2y + z, y + z)$  and the orderings  $<_{\text{lex}}, <_{\text{grlex}}$ .

**Solution.** □

For illustration, we present examples of several varieties defined by polynomials.

**12.D.2. Curves in the affine plane  $\mathbb{R}^2$ .** Every non-zero polynomial  $f(x, y)$  in two variables defines a “curve” in  $\mathbb{R}^2$  by the equation  $f(x, y) = 0$ . Thus, it is the set of zero points of a polynomial  $f$ , and it will be denoted  $K = \mathcal{V}(f)$ . You can derive that if  $f = f_1 \dots f_k$ , then  $\mathcal{V}(f) = \mathcal{V}(f_1) \cup \dots \cup \mathcal{V}(f_k)$ .

The subsequent pictures depict examples of such curves.

**12.D.3.** Using your favorite software, draw the curve given by the equation  $x^3 + x^2 - y^2 = 0$  in the plane.

**Solution.** See picture 1. □

AXIOMS OF CLASSICAL PROPOSITIONAL CALCULUS

**Definition.** *Axioms* of classical propositional calculus are:

- (kvp1)  $A \rightarrow (B \rightarrow A)$ ,
- (kvp2)  $(A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$ ,
- (kvp3)  $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$ ,
- (kvp4)  $\neg\neg A \rightarrow A$ ,

where  $A, B, C$  denote statements.

Axiom (kvp1) can be interpreted as follows: if statement  $A$  is valid, then we don’t need to know where it follows from; a valid statement follows from anything.

Axiom (kvp2) can be read as follows: if statement  $B$  follows from statement  $A$ , then if the negation of statement  $B$  also follows from statement  $A$ , statement  $A$  must necessarily not be valid. In other words, an implication is valid if its antecedent is not valid. In other words, a sufficient (but not necessary!) condition for the validity of an implication is the invalidity of the antecedent.

Axiom (kvp3) is called the axiom of deduction, and the reason for this name will become clear later. Axiom (kvp3), together with axiom (kvp1), expresses the transitivity of implication. However, this might not be immediately evident (if it is, I apologize to the reader for underestimating them). Therefore, we will also address the transitivity of implication later.

Axiom (kvp4) simply states that when it’s not true that statement  $A$  is not valid, then statement  $A$  is valid.

The selection of axioms for classical propositional calculus was arbitrary, and nothing forced us to choose them. However, as we will see later, this selection is very useful. Although it’s not the only possible one.

We can now construct statements and know four valid statements. What remains is to clarify how to derive new valid statements from valid ones. For this purpose, we will use the rule (1).

**Definition.** The *modus ponens inference rule* is of the form:

$$\frac{A \rightarrow B \quad A}{B}.$$

This rule states that if an implication is valid and its antecedent is valid, then its consequent is also valid. In other words, from the implication, we *deduce* the valid antecedent, and the valid consequent remains. Hence the name of this rule, *the method of detachment*.

We can summarize the first part of our considerations: The propositional logic system consists of three components. Countably infinite *alphabets* together with finitely many *syntactic rules* tell us how to construct *words* and *sentences*, i.e., *statements*, from individual *letters* of the alphabet. Further, a finite number of *axioms*, i.e., statements that we consider valid. The third component consists of a finite number of *inference rules*.

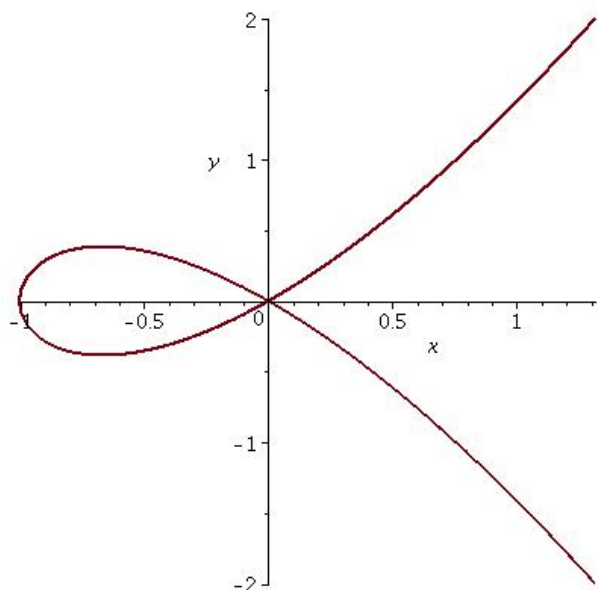


FIGURE 1.  $\mathfrak{V}(x^3 + x^2 - y^2)$

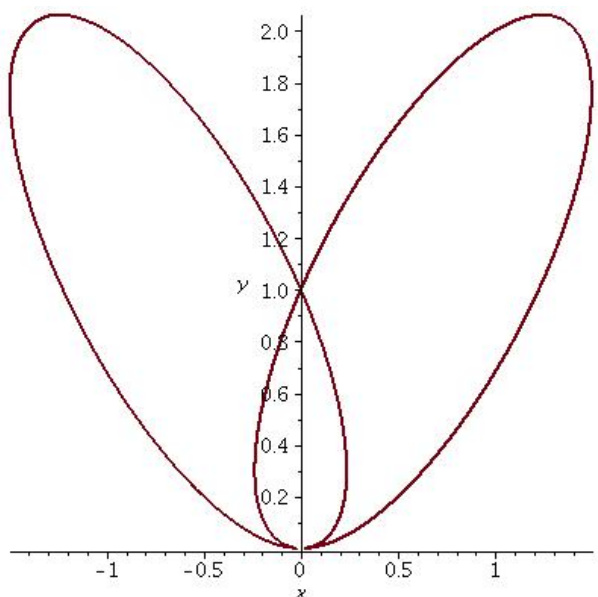


FIGURE 2.  $\mathfrak{V}(2x^4 - 3x^2y + y^2 - 2y^3 + y^4)$

**12.D.4.** Using your favorite software, draw the curve given by the equation  $2x^4 - 3x^2y + y^2 - 2y^3 + y^4 = 0$  in the plane.

**Solution.** See picture 2. □

We can also attempt to defined curves by equations  $x = f(t), y = g(t)$ , where  $f, g \in \mathbb{R}[t]$ . In that case, the curve is defined as a “polynomial inclusion” of the real line into the plane.

**12.D.5.** Parametrize the curve (variety)  $\mathfrak{V}(x^3 + x^2 - y^2)$ .

The specific choice of the alphabet with rules, axioms, and inference rules assigns some attribute to propositional calculus. In this case, that attribute is “classical.”

So far, we have introduced the primitive concepts of propositional calculus. It is already evident that the subject of (propositional) logic is the validity or invalidity of correctly constructed statements and logical deduction. Logic, in this sense, is a part of mathematics and has nothing to do with psychology or “rules of proper thinking.”

#### INFERENCE AND PROOFS

**Definition.** *Deriving a statement  $A$  from statements  $A_1, A_2, \dots, A_n$  is a finite sequence of statements, where each member is either an axiom, one of the statements  $A_1, A_2, \dots, A_n$ , or a statement derived from the previous members of the sequence using an inference rule. The last member of the derivation is the statement  $A$ .*

If there exists a derivation of statement  $A$  from statements  $A_1, A_2, \dots, A_n$ , we say that statement  $A$  can be derived from statements  $A_1, A_2, \dots, A_n$ , and we write

$$\{A_1, A_2, \dots, A_n\} \vdash A.$$

The statements in the set  $\{A_1, A_2, \dots, A_n\}$  are called the *premises* of the derivation, and statement  $A$  is called the *conclusion*.

If the premises of the derivation are valid statements, then the conclusion is also valid.

**Examples** that can be considered **claims**.

1) Transitivity of implication in classical propositional calculus

$$\{A \rightarrow B, B \rightarrow C\} \vdash A \rightarrow C.$$

We write down the derivation in a table. In the first column is the ordinal number of the statement in the derivation, in the second column is the corresponding statement, and in the third column is the justification for why the statement is included in the derivation. Typically, the justification refers to an axiom or refers to the lines from which the statement is derived using the modus ponens rule. Remember that uppercase letters are, in fact, parameters and can be replaced by any statement.

1.	A → B	premise
2.	B → C	premise
3.	(B → C) → (A → (B → C))	(kvp1)
4.	A → (B → C)	2., 3.
5.	(A → (B → C)) → ((A → B) → (B → C))	(kvp3)
6.	(A → B) → (B → C)	4., 5.
7.	A → C	1., 6.

We used exactly two axioms (kvp1) and (kvp3) in the derivation.

2) The Law of Excluded Middle

$$\{A, \neg A\} \vdash B,$$

also known as *ex impossibili sequitur quodlibet* (from nonsense, anything follows).



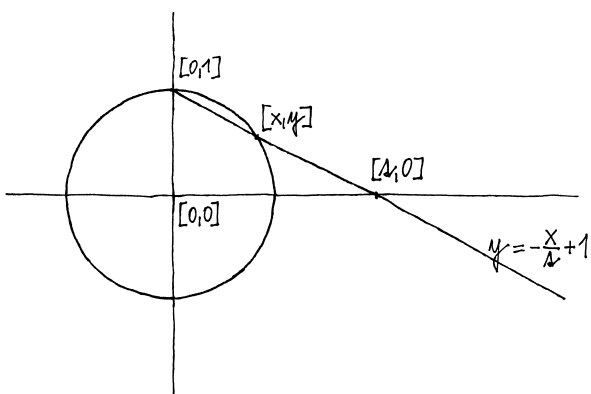
**Solution.** The parametrization can be found by computing intersections of lines  $y = tx$  with the given curve, i. e., we parametrize by the tangent of these lines. Technically, this means that we substitute  $tx$  for  $y$  and express  $x$  in terms of  $t$  from the equation:

$$x^3 + x^2 - t^2x^2 = x^2(x + 1 - t) \implies x = t - 1 \vee x = 0.$$

Then,  $y = t^2(t - 1)$ , or for  $x = 0$ , the only satisfying point on the curve is  $y = 0$ . The point  $[0, 0]$  can be obtained by choosing  $t = 1$  in the mentioned parametrization, so it suffices to consider only this parametrization.  $\square$

We obtain more curves if we consider quotients of polynomials in the parametrization, i. e.,  $f = \frac{f_1}{f_2}, g = \frac{g_1}{g_2}$ . Then, we talk about a rational parametrization.

**12.D.6.** Derive the parametrization of a circle using stereographic projection (see the picture).



**Solution.** Substituting the equation of the line  $y = \frac{x}{t} + 1$  into the equation of the circle, we get the equation

$$x^2 + \left( \frac{x^2}{t^2} - \frac{2x}{t} + 1 \right) = 1,$$

with solution  $x = 0$  or the parametric expression

$$x = \frac{2t}{1 + t^2}, y = \frac{t^2 - 1}{1 + t^2},$$

which does not include the point  $[0, 1]$ , however.  $\square$

**Remark.** Note that in this case, the inclusion of the real line gives only “almost all points” of the parametrized variety, since one of them (i. e., the point from which we project) is not reachable for any value of the parameter  $t$ . This is not our fault – it follows from different topological properties of the line and the circle that there exists no global parametrization.

**Remark.** Since  $\mathbb{R}$  is not an algebraically closed field, we have problems with existence of roots of polynomials. As a result, a mere perturbation of coefficients of the defining equation may drastically change the resulting variety. It is possible to work with complex polynomials  $\mathbb{C}[x, y]$  and with the subsets they define in  $\mathbb{C}^2$ . We need not be scared by that; on the contrary, our originally real curves are contained in their “complexifications” (real polynomials are simply viewed as complex which happen to have real coefficients), and we just

1. $A$	premise
2. $\neg A$	premise
3. $A \rightarrow (\neg B \rightarrow A)$	(kvp1)
4. $\neg B \rightarrow A$	1., 3.
5. $(\neg B \rightarrow A) \rightarrow ((\neg B \rightarrow \neg A) \rightarrow \neg\neg B)$	(kvp2)
6. $(\neg B \rightarrow \neg A) \rightarrow \neg\neg B$	4., 5.
7. $\neg A \rightarrow (\neg B \rightarrow \neg A)$	(kvp1)
8. $\neg B \rightarrow \neg A$	2., 7.
9. $\neg\neg B$	6., 8.
10. $\neg\neg B \rightarrow B$	(kvp4)
11. $B$	9., 10.

In these examples, we can see that it is easy to follow the derivation process. However, the construction of derivations can be very challenging. (This was not the case with the presented derivations.)

DEDUCTION AND PROOFS

**Definition.** Proof of statement  $A$  is a derivation from the empty set of premises. If there exists a proof of statement  $A$ , we say that this statement is *provable* and write  $\vdash A$ .

$A$  theorem is a provable statement.

**Example.** Theorem [1]:  $\vdash A \rightarrow A$ .

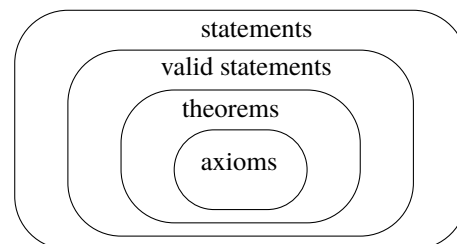
$A \rightarrow (A \rightarrow A)$	(kvp1)
$A \rightarrow ((A \rightarrow A) \rightarrow A)$	(kvp1)
$(A \rightarrow ((A \rightarrow A) \rightarrow A)) \rightarrow ((A \rightarrow (A \rightarrow A)) \rightarrow (A \rightarrow A))$	(kvp3)
$(A \rightarrow (A \rightarrow A)) \rightarrow (A \rightarrow A)$	2., 3.
$A \rightarrow A$	1., 4.

Derivations and proofs can be shortened by using previously derived statements or proven theorems. In the initial definition, we add that in the derivation of statement  $A$  from statements  $A_1, A_2, \dots, A_n$ , a theorem or a statement previously derived from the same premises can be included.

**Example.**  $\{A\} \vdash \neg\neg A$ .

1. $A$	premise
2. $A \rightarrow (\neg A \rightarrow A)$	(kvp1)
3. $\neg A \rightarrow A$	1., 2.
4. $(\neg A \rightarrow A) \rightarrow ((\neg A \rightarrow \neg A) \rightarrow \neg\neg A)$	(kvp2)
5. $(\neg A \rightarrow \neg A) \rightarrow \neg\neg A$	3., 4.
6. $\neg A \rightarrow \neg A$	Theorem [1]
7. $\neg\neg A$	5., 6.

A statement that can be derived from valid statements is also valid. Similarly, a theorem is a valid statement. In the propositional logic system, we naturally obtain an inclusion (hierarchy) of statements:



obtain richer tools for description of their properties (imaginary tangent lines, etc.).

Moreover, we are missing “improper points”. For instance, when parametrizing a circle, we can describe the missing points as the image of the only improper point of the real line, i. e. the point “at infinity”. These problems can be best avoided by working in the so-called projective extension of the (real or complex) plane.

The projective extension is advantageous to use in various problems, we will also use its application when defining the group operation on the points of an elliptic curve (see J).

**12.D.7. (The complex circle).** Consider the sets of points  $X^\varepsilon = \mathcal{V}(z_1^2 + z_2^2 - \varepsilon) \subseteq \mathbb{C}^2$  for any  $\varepsilon \in \mathbb{R} \setminus \{0\}$ . The corresponding real curves are

$$X_{\mathbb{R}}^\varepsilon = X^\varepsilon \cap \mathbb{R}^2 = \begin{cases} \text{circle with radius } \sqrt{\varepsilon} & \varepsilon > 0, \\ \emptyset & \varepsilon < 0. \end{cases}$$

We will write  $z_j = x_j + iy_j = x_j + \sqrt{-1}y_j$ . Therefore,  $X^\varepsilon$  is given as a subset in  $\mathbb{R}^4$  by a system of two real equations:

$$\begin{aligned} \operatorname{Re}(z_1^2 + z_2^2 - \varepsilon) &= x_1^2 + x_2^2 - y_1^2 - y_2^2 - \varepsilon = 0, \\ \operatorname{Im}(z_1^2 + z_2^2 - \varepsilon) &= 2(x_1y_1 + x_2y_2) = 0. \end{aligned}$$

Thus, we can assume that  $X^\varepsilon$  will be a “two-dimensional surface” in  $\mathbb{R}^4$ . We will try to imagine it as a surface in  $\mathbb{R}^3$  in a suitable projection  $\mathbb{R}^4 \rightarrow \mathbb{R}^3$ . For this purpose, we choose the mapping

$$\varphi_+ : (x_1, x_2, y_1, y_2) \mapsto \left( x_1, x_2, \frac{x_1y_2 - x_2y_1}{\sqrt{x_1^2 + x_2^2}} \right)$$

Denote by  $V$  the subset of  $\mathbb{R}^4$  which is given by our second equation, i. e.,

$$V = \{(x_1, x_2, y_1, y_2); x_1y_1 + x_2y_2 = 0, (x_1, x_2) \neq (0, 0)\}.$$

The restriction of  $\varphi_+$  to  $V$  is invertible, and its inverse  $\psi_+$  is given by

$$\psi_+ : (u, v, w) \mapsto \left( u, v, -\frac{vw}{\sqrt{u^2 + v^2}}, \frac{uw}{\sqrt{u^2 + v^2}} \right).$$

Now, note that

$$\left( \frac{x_1y_2 - x_2y_1}{\sqrt{x_1^2 + x_2^2}} \right)^2 = y_1^2 + y_2^2,$$

and hence it follows that

$$\varphi_+(V \cap X^\varepsilon) = H^\varepsilon = \{(u, v, w); u^2 + v^2 - w^2 - |\varepsilon| = 0\}.$$

Now, we can compose the constructed mappings

$$\varphi_\varepsilon : X^\varepsilon \rightarrow V @ > \varphi_+ >> \mathbb{R}^3 \setminus \{(0, 0, 0)\} \supseteq H^\varepsilon,$$

and for every  $\varepsilon > 0$ , we get a bijection  $\varphi_\varepsilon : X^\varepsilon \rightarrow H^\varepsilon$ . The real part of this variety is the “thinnest circle” on the one-part rotational hyperboloid  $H^\varepsilon$ ; see the picture.

The relationship between theorems and valid statements is discussed in the following

**Deduction Theorem.**

$$\begin{aligned} \{A_1, A_2, \dots, A_{n-1}, A_n\} \vdash A \\ \text{if and only if} \\ \{A_1, A_2, \dots, A_{n-1}\} \vdash A_n \rightarrow A. \end{aligned}$$

**PROOF.** First, assume that  $\{A_1, \dots, A_n\} \vdash A$ , i.e., we have some derivation of statement  $A$  from statements  $A_1, \dots, A_n$ , and this derivation has  $m$  steps. By induction on  $m$ , we will show that for each statement  $B$  in the derivation, statement  $A_n \rightarrow B$  can be derived from the statements  $\{A_1, \dots, A_{n-1}\}$ ; in particular, this holds for the last statement in the derivation,  $B = A$ .

If  $m = 1$ , the derivation consists of a single statement  $A$ . Then,  $A$  is either an axiom or one of the statements  $A_1, \dots, A_n$ . If  $A = A_n$ , then according to Theorem [1],  $A_n \rightarrow A$ . If  $A \neq A_n$ , then statement  $A_n \rightarrow A$  can be derived from axiom (kvp1) rewritten as  $A \rightarrow (A_n \rightarrow A)$  using one application of the modus ponens rule.

Now, assume that the statement holds for some  $m$ , and we have a derivation of statement  $A$  from statements  $A_1, \dots, A_n$  with a length of  $m + 1$ . The last statement in this derivation,  $A$ , necessarily resulted from applying the modus ponens rule (because we have no other rules) to some statements of the form  $\tilde{A}$  and  $\tilde{A} \rightarrow A$ , which appeared among the first  $m$  steps of the derivation. Therefore, the derivations of each of the statements  $\tilde{A}$  and  $\tilde{A} \rightarrow A$  can have a length of at most  $m$ . According to the induction hypothesis, we have:

$$\{A_1, \dots, A_{n-1}\} \vdash A_n \rightarrow \tilde{A}, \quad \{A_1, \dots, A_{n-1}\} \vdash A_n \rightarrow (A_n \rightarrow \tilde{A}).$$

We can now obtain statement  $A_n \rightarrow A$  by applying the modus ponens rule twice to the statement

$$(A_n \rightarrow (\tilde{A} \rightarrow A)) \rightarrow ((A_n \rightarrow \tilde{A}) \rightarrow (A_n \rightarrow A)),$$

which is axiom (kvp3).

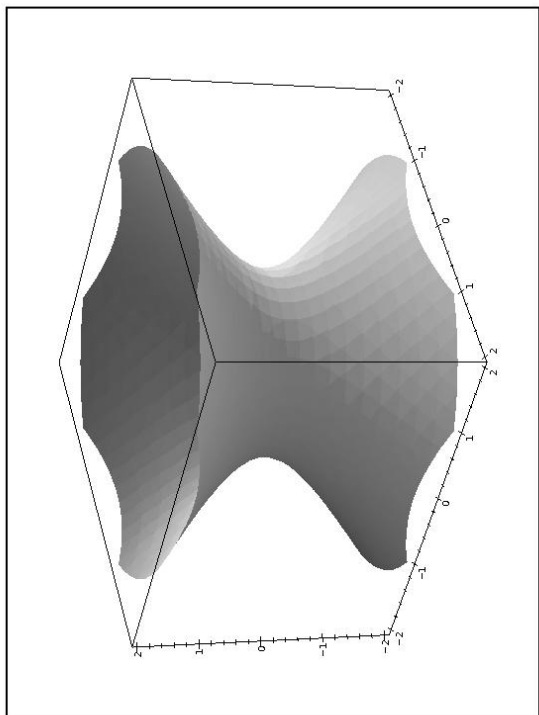
The necessity of the condition is proven. Its sufficiency is trivial since adding statement  $A_n$  to the premises  $A_1, \dots, A_{n-1}$  and using the modus ponens rule once yields  $\{A_1, \dots, A_{n-1}, A_n\} \vdash A$ .  $\square$

**Corollary.** Statement  $B$  can be derived from statement  $A$  if and only if statement  $A \rightarrow B$  is a theorem.

From this and the previous example, we can conclude that statement  $A \rightarrow \neg\neg A$  is a theorem. This statement is the converse of axiom (kvp4).

**Example.** 1) Theorem [2]:  $\vdash (A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$ .

We will show  $\{A \rightarrow B\} \vdash (B \rightarrow C) \rightarrow (A \rightarrow C)$ , and from this, using the Deduction Theorem, the statement follows. The abbreviation “TI.” in the sixth line of the derivation refers to the already proven transitivity of implication.



For  $\varepsilon < 0$ , we can repeat the above reasoning, merely interchanging  $x$  and  $y$  and the signs in the definition of  $\varphi_+$ :

$$\varphi_- : (x_1, x_2, y_1, y_2) \mapsto \left( -y_1, -y_2, \frac{-y_1x_2 + y_2x_1}{\sqrt{y_1^2 + y_2^2}} \right),$$

which changes the inversion  $\psi_-$

$$\psi_+ : (u, v, w) \mapsto \left( -\frac{vw}{\sqrt{u^2 + v^2}}, \frac{uw}{\sqrt{u^2 + v^2}}, -u, -v \right).$$

Now,  $H^\varepsilon$  is again a one-part rotational hyperboloid, but its real part is  $X_{\mathbb{R}}^\varepsilon = \emptyset$ .

In the complex case, we can observe that when continuously changing the coefficients, the resulting variety changes only a bit, except for certain “catastrophic” points, where a qualitative leap may occur. This is called the *principle of permanence*. In the real case, this principle does not hold at all.

**12.D.8. The projective extension of the line and the plane.**

The real projective space  $\mathbb{P}_1(\mathbb{R})$  is defined as the set of all directions in  $\mathbb{R}^2$ , i. e., its points are one-dimensional subspaces of  $\mathbb{R}^2$ .

The complex projective space  $\mathbb{P}_1(\mathbb{C})$  is defined as the set of all directions in  $\mathbb{C}^2$ , i. e., its points are one-dimensional subspaces of  $\mathbb{C}^2$ .

Similarly, the points of the real and complex two-dimensional projective spaces are defined as directions in  $\mathbb{R}^3$  and  $\mathbb{C}^3$ , respectively.

$A \rightarrow B$	premise
$(A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow B))$	(kvp1)
$(B \rightarrow C) \rightarrow (A \rightarrow B)$	1., 2.
$(B \rightarrow C) \rightarrow (A \rightarrow (B \rightarrow C))$	(kvp1)
$(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$	(kvp3)
$(B \rightarrow C) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$	4., 5., TI.
$((B \rightarrow C) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))) \rightarrow$	
$((B \rightarrow C) \rightarrow (A \rightarrow B)) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$	(kvp3)
$((B \rightarrow C) \rightarrow (A \rightarrow B)) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$	6., 7.
$(B \rightarrow C) \rightarrow (A \rightarrow C)$	3., 8.

2) Theorem [3]:  $\vdash (A \rightarrow (B \rightarrow C)) \rightarrow (B \rightarrow (A \rightarrow C))$ .

We will show  $\{A \rightarrow (B \rightarrow C), B, A\} \vdash C$ :

- |    |                                   |         |
|----|-----------------------------------|---------|
| 1. | $A \rightarrow (B \rightarrow C)$ | premise |
| 2. | $A$                               | premise |
| 3. | $B \rightarrow C$                 | 1., 2.  |
| 4. | $B$                               | premise |
| 5. | $C$                               | 3., 4.  |

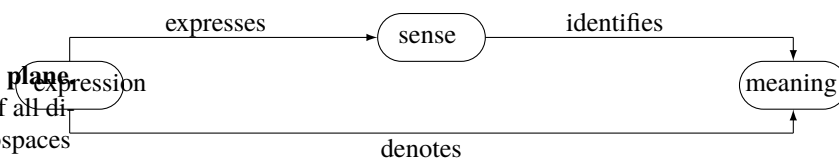
Now we use the Deduction Theorem three times to get

$$\{A \rightarrow (B \rightarrow C), B\} \vdash A \rightarrow C, \quad \{A \rightarrow (B \rightarrow C)\} \vdash B \rightarrow (A \rightarrow C), \\ \vdash (A \rightarrow (B \rightarrow C)) \rightarrow (B \rightarrow (A \rightarrow C)).$$

SEMANTICS OF CLASSICAL PROPOSITIONAL LOGIC

Classical propositional logic was introduced purely formally and axiomatically. An alphabet and syntax (rules for constructing “letters,” “words,” and “sentences”) were introduced, some statements (axioms) were declared valid, and a rule was specified for deriving other valid statements from valid ones. The idea that statements are or represent “declarative sentences that can be designated as true or false” and that we can somehow “read” propositional connectives is entirely unnecessary when creating a propositional logic system.

Now we will “fill the form with content,” show “what it is about.” More precisely, we will give meanings to statements. Let’s start somewhat broadly: We somehow understand linguistic expressions; they express some “meaning” for us. This meaning helps us determine or identify a fact or reality in the real (or fictional, or virtual) world that the respective linguistic expression denotes. This fact or reality is the “meaning” of the linguistic expression. Schematically:



A classic example, mentioned by Gottlob Frege, is the linguistic expression “Morning Star.” Its sense is the “brightest star in the morning sky.” Its meaning in the real world is the planet Venus. The sense of another expression, “Evening Star,” is the “brightest star in the evening sky.” But its meaning is the same—Venus.

**E. Algebraic structures**

First of all, we practice general properties of operations and we find out what structures the known sets and operations actually are.

**12.E.1.** Decide about the following sets and operations what algebraic structures they are (groupoid, semigroup (with potential one-sided neutral elements), monoid, group):

- i) the set of all subsets of the integers with union,
- ii) the set of positive integers with the greatest common divisor as the binary operation,
- iii) the set of positive integers with the least common multiple as the binary operation,
- iv) the set of all 2-by-2 invertible matrices over  $\mathbb{R}$  with addition,
- v) the set of all 2-by-2 matrices over  $\mathbb{R}$  with multiplication,
- vi) the set of all 2-by-2 matrices over  $\mathbb{R}$  with subtraction,
- vii) the set of all 2-by-2 invertible matrices over  $\mathbb{Z}_2$  with multiplication,
- viii) the set  $\mathbb{Z}_6$  with multiplication (modulo 6),
- ix) the set  $\mathbb{Z}_7$  with multiplication (modulo 7).

Construct the table of the operation for the last-but-two structure.

**Solution.**

- i) a monoid (the empty set being neutral),
- ii) a semigroup (with no neutral elements),
- iii) a monoid (1 being neutral),
- iv) not even a groupoid (consider  $A+(-A)$  for an invertible matrix  $A$ ),
- v) a monoid,
- vi) a groupoid (not associative),
- vii) a group,
- viii) a monoid (the class  $[1]$  being neutral),
- ix) a monoid (the class  $[1]$  being neutral).

The group in vii) consists of the following elements:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

$$D = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, E = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, F = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

The table of the matrix multiplication looks as follows:

	A	B	C	D	E	F
A	A	B	C	D	E	F
B	B	A	E	F	C	D
C	C	D	A	B	F	E
D	D	C	F	E	A	B
E	E	F	B	A	D	C
F	F	E	D	C	B	A

Note that each row and column (disregarding the heading ones) contains each element exactly once (why is it so?). Thus, we do not have to calculate each product and instead we can play “sudoku” as soon as we have filled enough entries of the table. □

The example shows that different expressions can have the same sense; conversely, we can also encounter expressions that have no sense. Statements in classical propositional logic are represented by various expressions. However, the meaning of a statement will be exactly one of the possibilities “true” or “false.” These meanings are symbolically denoted as T and F, sometimes as P and N in Czech, alternatively as 1 and 0. We will use this last notation. Sometimes it will be appropriate to interpret the symbols “0” and “1” directly as numbers, between which an order and arithmetic operations are naturally defined.

It is reasonable to consider the negation of a statement to be true precisely when the original statement is false, and vice versa. From the modus ponens rule, it follows that if an implication is true, and its antecedent is also true, then the consequent must also be true. Axiom (kvp2) implies that if the antecedent of an implication is false, then the entire implication is true. These considerations allow us to formulate a definition.

**Definition.** An interpretation of the classical propositional logic system is a mapping  $\mathcal{I}$  from the set of statements to the set  $\{0, 1\}$  such that

- (i1)  $\mathcal{I}(\neg A) = 1$  if and only if  $\mathcal{I}(A) = 0$ ,
- (i2)  $\mathcal{I}(A \rightarrow B) = 0$  if and only if  $\mathcal{I}(A) = 1$  and simultaneously  $\mathcal{I}(B) = 0$ .

The set  $\{0, 1\}$  is called the *set of truth values*.

The interpretations of primitive statements  $\neg A$  and  $A \rightarrow B$  can be written in tables:

A	$\neg A$	A	B	$A \rightarrow B$
0	1	0	0	1
0	1	0	1	1
1	0	1	0	0
1	0	1	1	1

We will use the notation  $|A| = \mathcal{I}(A)$ , which is often found in logic textbooks. Then we can express the interpretation of primitive statements as follows:

$$|\neg A| = 1 - |A|, \quad |A \rightarrow B| = \max\{1 - |A|, |B|\} = \max\{|\neg A|, |B|\}.$$

From the definition of interpretation, it is clear that the truth value of any statement can be computed for any interpretation of atomic statements by a finite number of subtractions and searching for the greater of the two values.

**Example.** Let  $|a| = 1, |b| = 0$ . Then

$$\begin{aligned} |\neg((b \rightarrow a) \rightarrow \neg a) \rightarrow (b \rightarrow a)| &= \max\{1 - |\neg((b \rightarrow a) \rightarrow \neg a)|, |b \rightarrow a|\} \\ &= \max\{|\neg((b \rightarrow a) \rightarrow \neg a)|, \max\{1 - |b|, |a|\}\} = \\ &= \max\{\max(1 - |b \rightarrow a|, |\neg a|), 1\} = 1. \end{aligned}$$

This ordered computation is, of course, not very efficient. A much shorter evaluation would be achieved using a truth table.

□ **Definition.**

**12.E.2.** Let  $X$  be a set and  $\mathcal{P}(X)$  denote the set of all subsets of  $X$ . Decide whether the set  $\mathcal{P}(X)$  together with each of the following operations forms a groupoid, semigroup, monoid, group and whether the operation is commutative:

- i) set intersection,
- ii) set union,
- iii) set symmetric difference (xor).

**Solution.** If the set  $X$  is empty, then  $\mathcal{P}(X)$  together with any of the mentioned operations is the trivial (1-element) group. Otherwise:

- i) with the set intersection, the resulting structure is a commutative monoid,
- ii) with the set union, the resulting structure is a commutative monoid,
- iii) with the set xor, the resulting structure is a commutative group where the empty set is neutral and each element is self-inverse.  $A^{-1} = A$ .

**12.E.3.** Decide about the following sets and operations what algebraic structures they are (groupoid, semigroup, group), whether they have one-sided or two-sided neutral elements, and whether the operation is commutative:

- i) the set of all 3-by-3 invertible matrices over  $\mathbb{R}$  with addition,
- ii) the set of all 3-by-3 matrices over  $\mathbb{R}$  with multiplication,
- iii) the set of all 3-by-3 matrices over  $\mathbb{R}$  with addition,
- iv) the set of all 3-by-3 invertible matrices over  $\mathbb{Z}_2$  with multiplication,
- v)  $(\mathbb{Z}_9, +)$ ,
- vi)  $(\mathbb{Z}_9, \cdot)$ .



**12.E.4.** Decide about the following subsets  $G$  of the complex numbers what algebraic structures they form together with multiplication (groupoid, semigroup, group), and whether the operation is commutative:

- i)  $G = \{a + bi \mid a, b \in \mathbb{Z}\}$ ,
- ii)  $G = \{a + bi \mid a, b \in \mathbb{R}, a^2 + b^2 = 1\}$ ,
- iii)  $G = \{a + b \cdot \sqrt{5} \mid a, b \in \mathbb{Q}, a^2 + b^2 \neq 0\}$ .



**12.E.5.** Decide whether  $\mathbb{Z}$  together with the operation  $\heartsuit$  forms a groupoid, semigroup, monoid, group, and whether  $\heartsuit$  is commutative, provided it is defined by:

- i)  $a \heartsuit b = (a, b)$ ,
- ii)  $a \heartsuit b = a^{|b|}$ ,
- iii)  $a \heartsuit b = 2a + b$ ,
- iv)  $a \heartsuit b = |a|$ ,
- v)  $a \heartsuit b = a + b + a \cdot b$ ,
- vi)  $a \heartsuit b = a + b - a \cdot b$ ,
- vii)  $a \heartsuit b = a + (-1)^a b$ .



- We say that a statement  $A$  is a *tautology* if, for every interpretation of the statements from which  $A$  is composed,  $|A| = 1$ .
- We say that a statement  $A$  is a *contradiction* if, for every interpretation of the statements from which  $A$  is composed,  $|A| = 0$ .
- We say that a statement  $A$  follows from a set of statements  $\{A_1, A_2, \dots\}$  (finite or countable) if, for an interpretation of the statements  $A_1, A_2, \dots$  such that  $|A_1| = |A_2| = \dots = 1$ , it holds that  $|A| = 1$ .
- We say that a set of statements  $\{A_1, A_2, \dots\}$  is semantically consistent if there exists an interpretation such that  $|A_1| = |A_2| = \dots = 1$ .

A tautology follows from an empty set of statements.

**Examples.**

- (1) The statement  $A \rightarrow \neg\neg A$  is a tautology:  
 If  $|A| = 1$ , then  $|A \rightarrow \neg\neg A| = \max\{1 - |A|, |\neg\neg A|\} = \max\{0, 1 - (1 - 1)\} = 1$ ,  
 If  $|A| = 0$ , then  $|A \rightarrow \neg\neg A| = \max\{1 - |A|, |\neg\neg A|\} = \max\{1, 1 - (1 - 0)\} = 1$ .

The given statement is also a theorem in classical vector calculus.

- (2) The statements  $A \rightarrow B$  and  $B \rightarrow C$  are semantically consistent.

We construct a table

$A$	$B$	$C$	$A \rightarrow B$	$B \rightarrow C$
0	0	0	1	1
0	0	1	1	1
0	1	0	1	0
0	1	1	1	1
1	0	0	0	1
1	0	1	0	1
1	1	0	1	0
1	1	1	1	1

and from it, we can see that both statements have an interpretation of 1 in four cases, specifically in rows 1, 2, 4, and 8.

- (3) The statement  $A \rightarrow C$  follows from the set of statements  $\{A \rightarrow B, B \rightarrow C\}$ .

From the previous example, we already know that the given set of statements is semantically consistent. In the interpretation table, we only need to focus on the rows where both statements have a truth value of 1.

$A$	$B$	$C$	$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow C$
0	0	0	1	1	1
0	0	1	1	1	1
0	1	1	1	1	1
1	1	1	1	1	1

All truth values in the last column of the table are equal to 1. Recall that the statement  $A \rightarrow C$  is also derivable from the set of statements  $\{A \rightarrow B, B \rightarrow C\}$ ; this is the transitivity of implication.

**12.E.6.** In how many ways can we fill the following table so that  $(\{a, b, c\}, *)$  would be a

- i) groupoid
- ii) commutative groupoid
- iii) a groupoid with a neutral element
- iv) a monoid
- v) a group

*	a	b	c
a	c	b	a
b			b
c			

**Solution.**

- i)  $3^5$
- ii) 9
- iii) 9
- iv) 1
- v) 0

□

**12.E.7.** Find the number of groupoids on a given three-element set.

**Solution.** Since the set is given, it remains to define the binary operation. In a groupoid, there is no restriction except that the result of the operation must be an element of the underlying set. Thus, for any pair of elements, there are three possibilities for the result. By the product rule, this gives

$$3^{3 \cdot 3} = 19683$$

groupoids.

□

**12.E.8.** Decide whether the set  $G = (\mathbb{R} \setminus \{0\} \times \mathbb{R})$  together with the operation  $\Delta$  defined by  $(x, y)\Delta(u, v) = (xu, xv+y)$  for all  $(x, y), (u, v) \in G$  is a groupoid, semigroup, monoid, group, and whether  $\Delta$  is commutative.

○

**12.E.9.** Let  $\Omega$  be a set with the multiplication operation defined by  $ab = a$  for all  $a, b$  in  $\Omega$ . Prove that such multiplication is associative.

○

**Solution.** For  $a, b, c \in \Omega$  we have:

$$a(bc) = ab = a = ac = (ab)c$$

□

**12.E.10.** Suppose that  $abc = e$  for some  $a, b, c$  in a group  $G$ . Show that  $bca = e$  as well.

○

**Solution.** Set  $x = bc$ . Then  $ax = e$  i.e.  $x^{-1} = a$  since  $x$  is invertible and so  $xa = (bc)a = e$ .

□

### F. Groups

We begin with recalling permutations and their properties. We have already met permutations in chapter two, see ??, where we used them to define the determinant of a matrix.

**12.F.1.** For each of the following conditions, find all permutations  $\pi \in \mathbb{S}_7$  which satisfy it:

- i)  $\pi^4 = (1, 2, 3, 4, 5, 6, 7)$

There are four possible truth tables for unary logical connectors. One of them is the table for the negation interpretation. We will show that the remaining three possible tables also express interpretations of statements formed using the unary connector  $\neg$  and the binary connector  $\rightarrow$ .

A	♠A	A → A	A	♣A	¬(A → A)	A	♥A	A
0	1	1	0	1	0	0	1	0
1	1	1	1	1	0	1	1	1

The first connector is called *unary verum*, the second *unary falsum*, and the third *assertion*; the symbols used are not standard.

Similarly, any truth table for binary logical connectors is satisfied by a statement created using the connectors  $\neg$  and  $\rightarrow$ . We will demonstrate this with at least one example.

A	B	A ↔ B	¬((A → B) → ¬(B → A))
0	0	1	1
0	1	0	0
1	0	0	0
1	1	1	1

You probably know that the connector  $\leftrightarrow B$  is called *equivalence*.

#### BASIC PROPERTIES OF CLASSICAL PROPOSITIONAL LOGIC

First, let's emphasize that it's essential to distinguish between concepts that are *axiomatic*, i.e., introduced using axioms and rules of inference, and those that are *semantic*, i.e., introduced using interpretation. (And let's also note that the concepts referred to as axiomatic are commonly called syntactic, but that's misleading because it leads to mixing axioms with syntax in the narrower sense—rules for forming statements.)

However, we've already seen that both the axiomatics and semantics include concepts that are related to each other—a "theorem" corresponds to a "tautology," "derivation" corresponds to "entailment." But this is the case only in classical propositional logic; in other propositional logic systems, it may not hold.

We've already formulated the basic properties of classical propositional logic, or at least hinted at them. Now let's summarize them without proofs:

**Deduction.:** The concept of a "theorem" is defined using the concept of "derivation" and is thus reducible to this concept. Conversely, the deduction theorem implies that the concept of "derivation from a finite set of statements" is reducible to the concept of a "theorem." The set of theorems coincides with the set of derivable statements.

The semantic counterpart of the deduction theorem also holds:

Statement  $A$  follows from a set of statements  $\{A_1, \dots, A_{n-1}, A_n\}$  if and only if statement  $A_n \rightarrow A$  follows from the set of statements  $\{A_1, \dots, A_{n-1}\}$ .

- ii)  $\pi^2 = (1, 2, 3) \circ (4, 5, 6)$
- iii)  $\pi^2 = (1, 2, 3, 4)$

○

**12.F.2.** Find the signature (parity) of each of the following permutations:

- i)  $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & \dots & 3n-2 & 3n-1 & 3n \\ 2 & 3 & 1 & 5 & 6 & 4 & \dots & 3n-1 & 3n & 3n-2 \end{pmatrix},$
- ii)  $\begin{pmatrix} 1 & 2 & 3 & \dots & n & n+1 & n+2 & \dots & 2n \\ 2 & 4 & 6 & \dots & 2n & 1 & 3 & \dots & 2n-1 \end{pmatrix}.$

**Solution.** The parity of a permutation corresponds to the number of transpositions from which it is built or, equivalently, to the number of its inversions, see 2.2.2. The number of inversions can be read easily from the two-row representation of the permutation. For each number of the second row, we count the number of numbers that are less and lie more to the right than the current number. Thus, the first permutation is even (the signature is 1), and in the second case, the signature depends on  $n$  and is equal to  $(-1)^{\frac{n \cdot (n+1)}{2}}$ . □

**12.F.3.** Find all permutations  $\rho \in \mathbb{S}_9$  such that

$$[\rho \circ (1, 2, 3)]^2 \circ [\rho \circ (2, 3, 4)]^2 = (1, 2, 3, 4).$$

**Solution.** No such permutation exists, since the left-hand side is always an even permutation, while the right-hand side is an odd one. □

**12.F.4.** Find all permutations  $\rho \in \mathbb{S}_9$  such that

$$\rho^2 \circ (1, 2) \circ \rho^2 = (1, 2) \circ \rho^2 \circ (1, 2).$$

○

**12.F.5.** Consider the permutation  $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 6 & 5 & 7 & 1 & 2 & 4 \end{pmatrix}.$

Find the order of  $\sigma$  in the group  $(\mathbb{S}_7, \circ)$ , the inverse of  $\sigma$  and compute  $\sigma^{2013}$ . Show that  $\sigma$  does not commute with the transposition  $\tau = (2, 3)$ .

**Solution.**  $\sigma = (1, 3, 5) \circ (2, 6) \circ (4, 7)$ . Therefore, the order of  $\sigma$  is the least common multiple of the cycle lengths 3, 2, 2, which is 6. Furthermore,  $\sigma^{-1} = (1, 5, 3) \circ (2, 6) \circ (4, 7)$  and

$$\sigma^{2013} = (\sigma^3 35)^6 \circ \sigma^3 = \sigma^3 = (2, 6) \circ (4, 7).$$

Finally, we have  $\sigma \circ \tau = (1, 3, 6, 2, 5) \circ (4, 7)$ , but  $\tau \circ \sigma = (1, 2, 6, 3, 5) \circ (4, 7)$ . □

**12.F.6.** Find  $\sigma^{-1}$  and  $\sigma^{2013}$ , where

- (a)  $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 5 & 7 & 6 & 1 & 2 & 3 \end{pmatrix}$  in the symmetric group  $(\mathbb{S}_7, \circ)$ .
- (b)  $\sigma = [4]_{11}$  in the group  $(\mathbb{Z}_{11}^\times, \cdot)$ .

**Solution.** (a)  $\sigma = (1, 4, 6, 2, 5) \circ (3, 7)$ ,  $\sigma^{-1} = (1, 5, 2, 6, 4) \circ (3, 7)$ , since the order of  $(1, 4, 6, 2, 5)$  is 5 and the order of the transposition  $(3, 7)$  is 2, we get that the order of  $\sigma$  is the least

**Compactness.:** A trivial consequence of the definition of derivation is that a statement is derivable from any set of statements if and only if it is derivable from some of its finite subsets. (This follows from the fact that derivation is a finite sequence.) This also holds for entailment:

A statement follows from an infinite set of statements if and only if it follows from some of its finite subsets.

It should be noted that proving this statement is not trivial.

**Correctness.:** Classical propositional logic is *correct*, meaning that every theorem is a tautology. "What can be proven is true."

**Completeness.:** Classical propositional logic is *complete*, meaning that every tautology is a theorem. "What is true can be proven."

**Strong Correctness and Completeness.:** Classical propositional logic is *strongly correct*, meaning that whenever a statement is derivable from a set of statements, it follows from it. Classical propositional logic is also *strongly complete*, meaning that whenever a statement follows from a set of statements, it is derivable from it.

Therefore, a statement in classical propositional logic follows from a set of statements if and only if it is derivable from it.

**Denotational Saturation.:** An operator with any truth table can be expressed using the operators  $\neg$  and  $\rightarrow$ ; that is, for any  $n$ -ary operator defined by a truth table, there exists a statement constructed solely using  $\neg$  and  $\rightarrow$  that has the same truth table.

**Decidability.:** For any (correctly formulated) statement, it can be decided whether it is a tautology or not. For any statement, it can be decided whether it is a theorem or not.

This statement follows directly from the definition of interpretation. Every statement is composed of only finitely many parts, that is, only finitely many atomic statements. And then, it can be decided by finite computation or a finite truth table whether the value is 1 for all evaluations of atomic parts.

ALTERNATIVE FORMULATIONS OF PROPOSITIONAL LOGIC

The primitive operations of propositional logic are, by definition, negation  $\neg$  and implication  $\rightarrow$ . The first is unary, the second is binary. Using them, we introduce several other binary operations on the set of statements, i.e., propositional functors:

statement	definition	name	
common multiple of 2 and 5, which is 10, i. e., $\sigma^{10} = 1$	$A \wedge B$	$\neg(A \rightarrow \neg B)$	conjunction
Then,	$A \vee B$	$\neg A \rightarrow B$	[inclusive] disjunction
$\sigma^{2013} = (\sigma^{10})^{201} \circ \sigma^3 = \sigma^3 = (1, 2, 4, 5, 6) \circ (3, 7)$	$A \underline{\vee} B$	$(A \wedge \neg B) \vee (\neg A \wedge B)$	exclusive disjunction
.	$A   B$	$\neg(A \wedge B)$	Sheffer's function, incompatibility
(b) For the sake of simplicity, we will write only $k$ for the	$A \downarrow B$	$\neg(A \vee B)$	Peirce's function
residue class $[k]_{11}$ , $k \in \mathbb{Z}$ . Then,	$A \leftrightarrow B$	$(A \rightarrow B) \wedge (B \rightarrow A)$	equivalence

$$4^5 \equiv 1 \pmod{11} \Rightarrow \sigma^{-1} = 4^4 \equiv 3 \pmod{11}$$

$$\sigma^{2013} = 4^{2013} \equiv 4^3 \equiv 9 \pmod{11}.$$

□

**12.F.7.** Prove that every group whose number of elements is even contains a non-trivial (i. e., different from the identity) element which is self-inverse.

**Solution.** Since each element which is not self-inverse can be paired with its inverse, we see that there are an even number of elements which are not-self inverse. Thus, there remain an even number of elements which are self-inverse, and one of them is the identity, so there must be at least one more such element. □

**12.F.8.** Prove that there exists no non-commutative group of order 4.

**Solution.** By Lagrange's theorem (see 12.4.10), the non-trivial elements of a 4-element group are of order 2 or 4. If there is an element of order 4, then the group is cyclic, and thus commutative. So the only remaining case is that there are (besides the identity  $e$ ) three elements of order 2, call them  $a$ ,  $b$ ,  $c$ . We are going to show that we must have  $ab = c$ : It cannot be that  $ab = e$ , since the inverse of  $a$  is  $a$  itself, and not  $b$ . It cannot be that  $ab = a$ , since this would mean that  $b = e$ , and similarly, it cannot be that  $ab = b$ , since this would mean that  $a = e$ . Therefore, the only remaining possibility is that indeed  $ab = c$ , and it can be shown analogously that the product of any two non-trivial elements, regardless of the order, must be equal to the third one, so this group is commutative, too. Altogether, we have shown that there are exactly two groups of order 4, up to isomorphism. The latter is called the Klein group, and one instance of it is the group  $\mathbb{Z}_2 \times \mathbb{Z}_2$ . □

**12.F.9.** Show that there exists no non-commutative group of order 5.

**Solution.** By Lagrange's theorem (see 12.4.10), the non-trivial elements of a 5-element group are of order 5, so the group must be cyclic, and thus commutative. □

**Remark.** The same argumentation show that each group of prime order must be cyclic, and thus commutative. In particular, there are neither 2-element nor 3-element non-commutative groups. As we have shown (see 12.F.8), there is even no 4-element non-commutative group. Therefore, the smallest non-commutative group may be of order 6. As we have seen (see 12.E.1(vii)), this is indeed the case.

All the connectors introduced in this way are not strictly part of propositional logic; they are meta-symbols. However, this strict distinction is not significant for our purposes, and we can freely add the symbols  $\wedge$ ,  $\vee$ ,  $\underline{\vee}$ ,  $|$ ,  $\leftrightarrow$ , and  $\downarrow$  to the propositional alphabet.

The conjunction  $A \wedge B$  can be read as "A and B" or "A simultaneously with B" and so on.

The disjunction  $A \vee B$  can be read as "A or B." The grammatical connector "or" here has a similar meaning to the Latin "vel," indicating possible but not always necessary alternatives, so at least one of the statements  $A$ ,  $B$  is true. Similarly, in the Czech sentence: "Tonight I will go to the pub or study mathematics." I can study mathematics with my classmates in the pub. In contrast, the exclusive disjunction  $A \underline{\vee} B$ , which can be read as "either A or B," expresses "or" in an exclusionary sense, meaning that exactly one of the statements  $A$ ,  $B$  is true, as in the Czech sentence: "Tonight I will go to the pub or somewhere else." The Czech language has another meaning for the word "or," which is that at most one of the options will occur, for example, in the sentence: "Tonight I will sleep either in Brno or in New York." This possibility is formalized by Sheffer's connector, and the statement  $A | B$  is suitable to read as "A is incompatible with B."

**Statement**  $A \downarrow B$  can be read as "neither A nor B," and the equivalence  $A \leftrightarrow B$  can be read as "A if and only if B" or "A is a necessary and sufficient condition for B."

Furthermore, it should be noted that the exclusive disjunction  $A \underline{\vee} B$  is sometimes denoted as  $A \leftrightarrow B$ , and the negation of implication  $\neg(A \rightarrow B)$  is sometimes called "inhibition" and denoted as  $A \rightarrow B$ . Alternative names for the Peirce function are  $\nabla$ ,  $\cdot|$ , or NOR.

The semantics of these propositional connectors can be easily derived from their definitions. Let's summarize them in a truth table:

A	B	$A \wedge B$	$A \vee B$	$A \underline{\vee} B$	$A   B$	$A \downarrow B$	$A \leftrightarrow B$
0	0	0	0	0	1	1	1
0	1	0	1	1	1	0	0
1	0	0	1	1	1	0	0
1	1	1	1	0	0	0	1

As an exercise, you can verify some "standard" tautologies:



**12.F.10.** Prove that any group  $G$  where each element is self-inverse must be commutative.

**Solution.** Let  $a, b \in G$ . Since each of  $ba, b, a$  is assumed to be self-inverse, we get

$$ab = ab((ba)(ba)) = a(bb)aba = (aa)ba = ba.$$

□

**12.F.11.** Prove that every group  $G$  of order 6 is isomorphic to  $\mathbb{Z}_6$  or  $\mathbb{S}_3$ .

**Solution.**

By Lagrange's theorem (see 12.4.10), the non-trivial elements of a 6-element group are of order 2, 3, or 6. If there is an element of order 6, then  $G$  is cyclic, and thus isomorphic to  $\mathbb{Z}_6$ .

Therefore, assume from now on that the order of each non-trivial element is 2 or 3. Since an element  $a$  of order 3 is not self-inverse (we have  $a^{-1} = a^2$  since  $a \cdot a^2 = a^3 = 1$ ), we get from exercise 12.F.7 that there must be at least one element of order 2.

As we are going to show, there must also be an element of order 3. For the sake of contradiction, assume that each element of  $G$  is self-inverse, and let  $a \neq b$  be any two elements different from the identity  $e$ . The same argumentation as in 12.F.8 shows that the product  $ab$  cannot be any of  $e, a, b$ . Thus,  $H = \{e, a, b, ab\}$  is a 4-element subset of  $G$ . Thanks to the self-inverseness, we can see that  $H$  is closed under the operation, with the possible exception of  $b \cdot a, b \cdot ab$ , and  $ab \cdot a$ . However, we get from the above exercise that  $G$  is commutative, so that these 3 products also lie in  $H$ , and it follows that  $H$  is actually a subgroup of  $G$ . However, this contradicts theorem 12.4.10, by which a 6-element group cannot have a 4-element subgroup.

The only remaining case is that there is an element of order 2 (call it  $a$ ) as well as an element of order 3 (call it  $b$ ). Then,  $b^2$  is also of order 3 (and different from  $b$ ), so  $G$  contains the 4 elements  $e, a, b, b^2$ . Furthermore,  $G$  must also contain  $ab, ba, ab^2, b^2a$ , and by uniqueness of inverses, none of these is equal to  $e$ . Moreover, none of these may be equal to any of  $a, b, b^2$  (e. g. if we had  $a = ab$ , then multiplication by  $a^{-1}$  from the left yields  $e = b$ ; the other equalities can be refuted similarly). Since  $G$  contains only 6 elements, the set  $\{ab, ba, ab^2, b^2a\}$  has at most two. Again, we can have neither  $ab = ab^2$  nor  $ba = b^2a$ . If  $ab = ba$ , then  $(ab)^2 = a^2b^2 = b^2 \neq e$  and  $(ab)^3 = a^3b^3 = a \neq e$ , so the order of  $ab$  is greater than 3, which contradicts our assumption. Therefore, it must be that  $ab = b^2a$  and  $ba = a^2b$ , so that  $G$  is indeed isomorphic to  $\mathbb{S}_3$  ( $a$  corresponds to a transposition and  $b$  does to a cycle of length 3). This group can also be viewed as the group of symmetries of an equilateral triangle ( $a$  corresponds to a reflection and  $b$  does to a rotation by  $120^\circ$ ), see also 12.4.3.

We have discussed all possibilities, so the proof is finished. □

$A \vee \neg A, \neg(A \wedge \neg A)$	Law of excluded middle
$(A \wedge \neg A) \rightarrow B, \neg A \rightarrow (A \rightarrow B)$	Law of Duns Scotus
$((A \rightarrow B) \wedge A) \rightarrow B$	Modus ponens
$(\neg B \rightarrow \neg A) \rightarrow (A \rightarrow B)$	Reductio ad absurdum, indirect
$((A \wedge \neg B) \rightarrow (C \wedge \neg C)) \rightarrow (A \rightarrow B)$	Proof by contradiction
$((A \rightarrow B) \wedge (B \rightarrow C)) \rightarrow (A \rightarrow C)$	Transitivity of implication
$((A \wedge B) \rightarrow C) \rightarrow (A \rightarrow (B \rightarrow C))$	Deduction

We can declare other connectors, such as  $\neg, \wedge,$  and  $\vee,$  as primitive operations alongside  $\rightarrow$ . In this sense, Propositional Calculus can be treated as discussed in Section 11.48. For constructing propositional calculus, one of either  $\wedge$  or  $\vee$  alongside  $\neg$  is sufficient, as implication  $\rightarrow$  can be expressed using them:

$$A \rightarrow B = \neg A \vee B, \quad \text{or} \quad A \rightarrow B = \neg(A \wedge \neg B).$$

Axioms can then be rewritten using the respective connector. The inference rules will take the following form:

$$\frac{A \vee B}{\neg A} \quad \frac{A \vee B}{\neg B} \quad \text{or} \quad \frac{\neg(A \wedge B)}{A} \quad \frac{\neg(A \wedge B)}{B}$$

Each of the Sheffer and Peirce functions can serve as a single propositional connector. Specifically:

$$\neg A = A | A, \quad A \rightarrow B = A | (B | B),$$

and

$$\neg A = A \downarrow A, \quad A \rightarrow B = ((A \downarrow A) \downarrow B) \downarrow ((A \downarrow A) \downarrow B).$$

The inference rules can then be written as:

$$\frac{A | B}{A} \quad \frac{A | B}{B} \quad \text{and} \quad \frac{A \downarrow B}{A \downarrow A} \quad \frac{A \downarrow B}{B \downarrow B}$$

As previously mentioned, Propositional Calculus can be built using other connectors than  $\neg$  and  $\rightarrow$ . However, any combination of these connectors will suffice, as shown in Section 11.48. The construction of Propositional Calculus will only require one of  $\wedge$  or  $\vee$  in addition to negation, as implication can be expressed using them.

Therefore, the axioms for classical propositional calculus can be either:

$$\begin{aligned} \text{(kvp1)} \quad & A \rightarrow (B \rightarrow A) \\ \text{(kvp2')} \quad & (\neg A \rightarrow \neg B) \rightarrow ((\neg A \rightarrow B) \rightarrow A) \\ \text{(kvp3)} \quad & (A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)) \end{aligned}$$

Or, a single axiom known as the "Meredith's Axiom":

$$\left( \left( \left( (A \rightarrow B) \rightarrow (\neg C \rightarrow \neg D) \right) \rightarrow C \right) \rightarrow E \right) \rightarrow ((E \rightarrow A) \rightarrow (D \rightarrow A)).$$

On the other hand, if we consider  $\neg, \rightarrow, \wedge,$  and  $\vee$  as primitive connectors (i.e., not defining conjunction and disjunction in terms of implication and negation), we would need

**12.F.12.** Find all commutative groups of order 8 (up to isomorphism). Then, for each of the following groups, decide to which of the found ones it is isomorphic (the operation is always multiplication):

- $\mathbb{Z}_{15}^\times$ ,
- $\mathbb{Z}_{16}^\times$ ,
- $\mathbb{Z}_{17}^\times / \{[1], [-1] = [16], \cdot\}$ ,
- the complex roots of the polynomial  $z^8 - 1$ .

**Solution.**

By theorem 12.4.8, every commutative group is a product of cyclic groups. By 12.4.10, their orders divide 8. This means that there are only 3 possibilities:  $\mathbb{Z}_8$ ,  $\mathbb{Z}_2 \times \mathbb{Z}_4$ , and  $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ .

- The group  $\mathbb{Z}_{15}^\times$  contains the residue classes which are coprime to 15. There are  $\varphi(15) = (5 - 1)(3 - 1) = 8$  of them, so indeed  $|\mathbb{Z}_{15}^\times| = 8$ . In particular, these are 1, 2, 4, 7, 8, 11, 13, 14. Their orders are either 2 (for 4, 11, 14) or 4 (for 2, 7, 8, 13), which means that  $\mathbb{Z}_{15}^\times$  is isomorphic to  $\mathbb{Z}_2 \times \mathbb{Z}_4$ .
- $\mathbb{Z}_{16}^\times = \{1, 3, 5, 7, 9, 11, 13, 15\}$ . Again, this group contains 8 elements, and their orders are either 2 (for 7, 9, 15) or 4 (for 3, 5, 11, 13), which means that  $\mathbb{Z}_{16}^\times$  is also isomorphic to  $\mathbb{Z}_2 \times \mathbb{Z}_4$ .
- $\mathbb{Z}_{17}^\times = \{\pm 1, \pm 2, \dots, \pm 8\}$ . Thus, the quotient  $\mathbb{Z}_{17}^\times / (\pm 1) = \{1, 2, \dots, 8\}$  has 8 elements. We can easily calculate that the order of 3 is 8. Therefore, 3 generates the entire group, which means that  $\mathbb{Z}_{17}^\times / (\pm 1) \cong \mathbb{Z}_8$ .
- The complex roots of the polynomial  $z^8 - 1$  are  $e^{\frac{n\pi}{4}i}$ , where  $n = 1, 2, \dots, 8$ . Clearly, these form a cyclic group of order 8, isomorphic to  $\mathbb{Z}_8$ .

**12.F.13.** Let  $G$  be a commutative group and denote  $H = \{g \in G \mid g^2 = e\}$ , where  $e$  is the identity of  $G$ . Prove that  $H$  is a subgroup of  $G$ .

**Solution.** Clearly,  $e \in H$ . If  $a \in H$ , then we also have  $a^{-1} \in H$ , because  $a = a^{-1}$  (since  $a^2 = e$ ). Moreover, if  $b \in H$ , then  $(ab)^2 = a^2b^2 = e$  (this is where we use the commutativity of  $G$ ), which means that  $ab \in H$ . Thus,  $H$  is closed under the operation, and it is indeed a subgroup.  $\square$

**12.F.14.** Let  $\mathcal{GL}_n(\mathbb{R})$  denote the set of all  $n$ -by- $n$  regular matrices with real coefficients. Prove that  $G = \mathcal{GL}_2(\mathbb{R})$  with multiplication is a group and decide for each of the following subsets  $H$  of  $G$  whether it is a subgroup of  $G$ :

- i)  $H = \mathcal{GL}_2(\mathbb{Q})$ ,
- ii)  $H = \mathcal{GL}_2(\mathbb{Z})$ ,
- iii)  $H = \{A \in \mathcal{GL}_2(\mathbb{Z}) \mid |A| = 1\}$ ,
- iv)  $H = \left\{ \begin{pmatrix} 0 & a \\ a & b \end{pmatrix} \in G \mid a, b \in \mathbb{Q} \right\}$ ,
- v)  $H = \left\{ \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \in G \mid a \in \mathbb{Z} \right\}$ ,
- vi)  $H = \left\{ \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \in G \mid a \in \mathbb{Q} \right\}$ ,
- vii)  $H = \left\{ \begin{pmatrix} 0 & a \\ b & c \end{pmatrix} \in G \mid a, b, c \in \mathbb{R} \right\}$ ,

a more extensive system of axioms. Kleene introduced such axioms in 1967:

- (Kkvp1)  $A \rightarrow (B \rightarrow A)$
- (Kkvp2)  $((A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)))$
- (Kkvp3)  $(A \wedge B) \rightarrow A$
- (Kkvp4)  $(A \wedge B) \rightarrow B$
- (Kkvp5)  $A \rightarrow (B \rightarrow (A \wedge B))$
- (Kkvp6)  $A \rightarrow (A \vee B)$
- (Kkvp7)  $B \rightarrow (A \vee B)$
- (Kkvp8)  $(A \rightarrow C) \rightarrow ((B \rightarrow C) \rightarrow ((A \vee B) \rightarrow C))$
- (Kkvp9)  $(A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$
- (Kkvp10)  $\neg\neg A \rightarrow A$

Notably, axioms (Kkvp1), (Kkvp2), (Kkvp9), and (Kkvp10) are the same as axioms (kvp1), (kvp3), (kvp2), and (kvp4).

#### ALTERNATIVES TO CLASSICAL PROPOSITIONAL CALCULUS

Classical propositional calculus possesses beautiful properties, particularly decidability and strong completeness and correctness. This means that in classical propositional calculus, the axiom system essentially aligns with semantics. However, this can also pose a danger - one might blur the line between semantics and axiomatics and perceive all of mathematics as precise formal calculations.

One of the problems with classical propositional calculus is the primitive concept of  $\rightarrow$ . It is meant to correspond to the intuitive notion of implication. However, consider the Czech sentences: "If it is dry in the Sahara, then beer is brewed in the Czech Republic" (which is true due to the truth of the consequent) and "If the Illuminati rule the world, they caused a tsunami in Sri Lanka" (which is true due to the falsity of the antecedent). These are certainly not exemplary true statements. Sometimes, constructions like these, which are consequences of axiom (kvp1) and theorem  $\neg A \rightarrow (A \rightarrow B)$ , are referred to as "paradoxes of implication."

Another problem is posed by axiom (kvp4), which is semantically interpreted as implying that if a statement is not false, then it is true. This appears to be self-evident, but it conceals an unspoken assumption that there are no truth values other than "true" and "false." A statement could have another truth value, such as "possible," or it could have no truth value, expressing uncertainty or "I don't know." Accepting this possibility means relinquishing the "view from God's eye."

Alternative approaches to propositional calculus may involve "improving implication" or introducing multiple truth values. We will explore some significant cases.

**(1) Intuitionistic Propositional Calculus.** This arises through a slight modification of classical propositional calculus. Axiom (kvp4) is replaced by a weaker axiom:

- (ivp4)  $A \rightarrow (\neg A \rightarrow B)$ .

This axiom is weaker in the sense that it is a theorem of classical propositional calculus (which can be easily verified by showing that (ivp4) is a tautology of classical propositional

viii)  $H = \left\{ \begin{pmatrix} 1 & a \\ b & c \end{pmatrix} \in G \mid a, b, c \in \mathbb{R} \right\}$ .

**12.F.15.**

- i) Decide whether the set  $H = \{a \in \mathbb{R}^* \mid a^2 \in \mathbb{Q}\}$  is a subgroup of the group  $(\mathbb{R}^*, \cdot)$
- ii) Decide whether the set  $H = \{a \in \mathbb{R} \mid a^2 \in \mathbb{Q}\}$  is a subgroup of the group  $(\mathbb{R}, +)$

**12.F.16.** Find all positive integers  $m \neq 5$  such that the group  $\mathbb{Z}_m^\times$  is isomorphic to  $\mathbb{Z}_5^\times$ .

**12.F.17.** How many cycles of length  $p$  ( $1 < p \leq n$ ) are there in  $\mathbb{S}_n$ ?

**Solution.** The elements of the cycle (i. e., the non-fixed points of the permutation) can be selected in  $\binom{n}{p}$  ways. Now, without loss of generality, we can proclaim one of the  $p$  elements to be the first element in the cycle representation (for instance the least one, if we are working with numbers). This element can be mapped to any of the  $p - 1$  remaining elements, that one can be mapped to any of the  $p - 2$  remaining elements, etc. Altogether, we get by the product rule that there are  $\binom{n}{p \cdot (p-1)!}$  cycles of length  $p$ .  $\square$

**12.F.18.** Let  $G$  be the set of real-valued matrices with zeros above the diagonal and ones on it. Prove that  $G$  with matrix multiplication forms a group, i. e., a subgroup in  $\mathcal{GL}(3, \mathbb{R})$ , and find the center of  $G$  (i. e., the subgroup defined by  $Z(G) = \{z \in G \mid \forall g \in G : zg = gz\}$ ).

**Solution.** We can either verify all the group axioms or make use of the known fact that  $\mathcal{GL}(3, \mathbb{R})$  is a group, and we verify only that  $G$  is closed with respect to multiplication and inverses. Clearly, the neutral element (the identity matrix) lies in  $G$ .

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ a_1 & 1 & 0 \\ b_1 & c_1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a + a_1 & 1 & 0 \\ b + ca_1 + b_1 & c + c_1 & 1 \end{pmatrix} \in G,$$

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ -b + ac & -c & 1 \end{pmatrix} \in G.$$

It follows from the form of the products in  $G$  that the center contains precisely the matrices of the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix}.$$

**12.F.19.** For any subset  $X \subseteq G$ , we define its centralizer as  $C_G(X) = \{y \in G \mid xy = yx, \text{ for all } x \in X\}$ . Prove that if  $X \subseteq Y$ , then  $C_G(Y) \subseteq C_G(X)$ . Further, prove that  $X \subseteq C_G(C_G(X))$  and  $C_G(X) = C_G(C_G(C_G(X)))$ .

**Solution.** The first proposition is clear: The elements of  $G$  which commute with everything from  $Y$  also commute with everything from  $X$ . We have from the definition that

calculus), but (kvp4) is not a theorem of intuitionistic propositional calculus. This means that anything provable or derivable in intuitionistic propositional calculus can also be proven in classical propositional calculus. All the theorems of classical propositional calculus proven without using axiom (kvp4) are also theorems of intuitionistic propositional calculus.

In intuitionistic propositional calculus, the Deduction Theorem also holds. In its proof, we did not use axiom (kvp4).

The connection between classical and intuitionistic propositional calculus is further demonstrated by the following

**Reflection Theorem.** If statement  $A$  is derivable from statements  $A_1, \dots, A_n$  in classical propositional calculus, then it is also derivable from statements  $\neg\neg A_1, \dots, \neg\neg A_n$  in intuitionistic propositional calculus.

Classical reasoning is thus mirrored in intuitionistic calculus through double negation. In this sense, intuitionistic propositional calculus is stronger than classical. These observations lead to the conclusion that the  $\rightarrow$  operator has a different meaning in classical and intuitionistic logic, even though it is denoted by the same symbol. (The meaning of an expression can depend on the context in which it is used.)

Therefore, the semantics of intuitionistic logic is different from classical logic. It cannot be defined using truth tables; it is introduced using the concept of possible worlds. However, intuitionistic propositional calculus is just as "beautiful" as classical, being strongly correct and complete, compact, and decidable.

**(2) Lukasiewicz's Three-Valued Propositional Calculus.** The set of truth values includes not only 0 and 1 but also the value  $x$ , which can be interpreted as "I don't know." Their semantics are defined by truth tables as follows:

$A$	$\neg A$
0	1
$x$	$x$
1	0

$A \rightarrow B$	$B$
	0 $x$ 1
$A$	0 1 1 1
	$x$ $x$ $x$ 1
	1 0 $x$ 1

$A \wedge B$	$B$
	0 $x$ 1
$A$	0 0 0 0
	$x$ 0 $x$ $x$
	1 0 $x$ 1

$A \vee B$	$B$
	0 $x$ 1
$A$	0 0 $x$ 1
	$x$ $x$ $x$ 1
	1 1 1 1

The axioms of this propositional calculus are:

- (I<sub>3</sub>vp1)  $A \rightarrow (B \rightarrow A)$ ,
- (I<sub>3</sub>vp2)  $(A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$ ,
- (I<sub>3</sub>vp3)  $(\neg A \rightarrow \neg B) \rightarrow (B \rightarrow A)$ ,
- (I<sub>3</sub>vp4)  $((A \rightarrow \neg A) \rightarrow A) \rightarrow A$ .

The modus ponens rule is used for inference.

Lukasiewicz's three-valued propositional calculus is decidable, compact, strongly correct, and complete. However, the Deduction Theorem does not hold in it.

**(3) Lukasiewicz's fuzzy propositional calculus.** The set of truth values is a closed interval  $[0, 1]$ . The values 0 and

$C_G(C_G(X)) = \{y \in G \mid xy = yx, \forall x \in C_G(X)\}$ , and this is in particular satisfied by the elements  $y \in X$ . The last statement follows simply using the two above. Substituting  $X := C_G(X)$  into the second one, we get  $C_G(X) \subseteq C_G(C_G(C_G(X)))$ , and applying the first one to the second one, we obtain  $C_G(X) \supseteq C_G(C_G(C_G(X)))$ .  $\square$

**12.F.20.** Suppose that a group  $G$  has a non-trivial subgroup  $H$  which is contained in every non-trivial subgroup of  $G$ . Prove that  $H$  is contained in the center of  $G$ .

**Solution.** For each  $g \in G$ , the centralizer  $C_G(g) = \{x \in G \mid xg = gx\}$  is a non-trivial subgroup, since  $g \in C_G(g)$  and  $C_G(e) = G$ . Thus, the group  $H$  is contained in every  $C_G(g)$ . Therefore, it is contained in their intersection (over all  $g \in G$ ), which is exactly the center of  $G$ .  $\square$

**12.F.21.** Let  $G$  be a finite group. The conjugation class for  $a \in G$  is the set

$$Cl(a) = \{xax^{-1} \mid x \in G\}.$$

Prove that:

- i) the set of conjugation classes of all elements of  $G$  is a partition of  $G$ ,
- ii) the size of each conjugation class divides the order of  $G$ ,
- iii) if  $G$  has only two conjugation classes, then its order is 2.

**Solution.** (i) It suffices to show that we have for any  $a, b \in G$  that either  $Cl(a) = Cl(b)$  or  $Cl(a) \cap Cl(b) = \emptyset$ . Thus, assume that the intersection of  $Cl(a)$  and  $Cl(b)$  is non-empty. Then, by definition, there are  $x, y \in G$  such that  $xax^{-1} = yby^{-1}$ . Multiplying this equality by  $y^{-1}$  from the left and by  $y$  from the right leads to  $y^{-1}xax^{-1}y = b$ . However,  $(y^{-1}x)^{-1} = x^{-1}y$ , which means that  $b$  is of the form  $zaz^{-1}$  for  $z = y^{-1}x$  and thus lies in  $Cl(a)$ . Analogously, we get  $a \in Cl(b)$ , so that both conjugation classes coincide.

(ii) Note that the elements of  $Cl(a)$  are in one-to-one correspondence with the cosets corresponding to the centralizer  $C_G(a) = \{x \in G \mid xax^{-1} = a\}$ . Indeed, if elements  $b$  and  $c$  lie in the same coset (i. e., they satisfy  $b = cz$  for some  $z \in C_G(a)$ ), then

$$bab^{-1} = cza(cz)^{-1} = czaz^{-1}c^{-1} = czz^{-1}ac^{-1} = cac^{-1}.$$

By 10.2.1, we have  $|G| = |C_G(a)| \cdot |G/C_G(a)|$ , which means that  $|Cl(a)| = |G/C_G(a)|$  divides  $|G|$ .

(iii) The neutral element always forms its own conjugation class  $Cl(e) = \{e\}$ . Therefore, if there are only two conjugation classes, then all the other elements  $a \neq e$  must lie in one class. Thus, its size is  $|G| - 1$ , and by (ii), this integer must divide  $|G|$ , which means that  $|G| = 2$ .  $\square$

**12.F.22.** Let  $G$  be a commutative group. Suppose that the order  $r$  of an element  $a \in G$  and the order  $s$  of an element  $b \in G$  are coprime. Prove that the order of  $ab$  is  $rs$ .

**Solution.** We have  $(ab)^{rs} = a^{rs}b^{rs} = (a^r)^s(b^s)^r = e^s e^r = e$ , so the order is at most  $rs$ . For the sake of contradiction, assume that  $(ab)^q = e$  for some  $q < rs$ . Since  $q$  is less than the least common multiple of  $r$  and  $s$  (recall that  $r, s$  are

1 represent truth and falsehood, just like in classical propositional calculus. The values in between can be seen as the degree of certainty about the statement. The interpretation is governed by the rules:

$$|\neg A| = 1 - |A|, \quad |A \rightarrow B| = \max\{1 - |A|, |B|\} = \max\{|\neg A|, |B|\},$$

$$|A \wedge B| = \min\{|A|, |B|\}, \quad |A \vee B| = \max\{|A|, |B|\}.$$

The first two rules are the same as in classical propositional calculus. The degree of certainty of a statement cannot be interpreted as its probability, even when the statements are about random events. In such cases, the conjunction  $\wedge$  would express the simultaneous occurrence of events. The degree of certainty is calculated directly from the degrees of certainty of individual events, not their probabilities, as it also depends on the stochastic dependence or independence of the relevant events.

The axioms of this propositional calculus are:

- (1<sub>F</sub>vp1)  $A \rightarrow (B \rightarrow A)$ ,
- (1<sub>F</sub>vp2)  $(A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$ ,
- (1<sub>F</sub>vp3)  $(\neg A \rightarrow \neg B) \rightarrow (B \rightarrow A)$ ,
- (1<sub>F</sub>vp4)  $((A \rightarrow B) \rightarrow B) \rightarrow ((B \rightarrow A) \rightarrow A)$ .

The modus ponens rule is used for inference. The first three axioms of fuzzy propositional calculus are the same as the axioms of the three-valued propositional calculus.

Fuzzy propositional calculus is decidable and compact, but it is not strongly correct or complete. The Deduction Theorem does not hold in it.

#### (4) Paraconsistent four-valued propositional calculus.

We previously interpreted the truth value  $x$  in Lukasiewicz's three-valued propositional calculus as "I don't know." However, we can also interpret it as meaning that a given statement has no truth value, i.e., "neither true nor false." Now, to the three truth values 0, 1, and  $x$ , we add a fourth value, denoted as  $\mathfrak{g}$ , and interpret it as "simultaneously true and false." This may sound nonsensical at first, but these introduced truth values can have meaning when "seeking the truth" in the real world. For example, imagine that we have witness statements about a situation that can be expressed by a statement (declarative sentence). If the witnesses agree that the situation occurred, we assign a truth value of 1 to the statement. If they agree that the situation did not occur, we assign a value of 0. However, it's possible that witness statements may differ. In such a case, we assign the truth value  $\mathfrak{g}$  to the statement. And if there is no testimony at all, the statement will have the value  $x$ .

The value  $x$  is also referred to as a "truth-value gap," while the value  $\mathfrak{g}$  is called a "truth-value glut." We can define the interpretation as an ordered pair  $(\mathcal{I}_1, \mathcal{I}_2)$  of interpretations of atomic statements in Lukasiewicz's three-valued logic, with two witnesses interpreting the statement. Each of them can respond with "yes," "no," or "I don't know." To these pairs of interpretations, we assign truth values in the interpretation  $\mathcal{I}$  according to the table:

coprime), at least one of them does not divide  $q$ . Assume that it is  $r$  (the other case can be refuted analogously). Taking the  $s$ -th power of the equality  $(ab)^q = e$ , we get  $e = ((ab)^q)^s = (ab)^{qs} = a^{qs}b^{qs} = a^{qs}(b^s)^q = a^{qs}e^q = a^{qs}$ . Since  $r$  does not divide  $q$  and is coprime to  $s$ , we get that  $r$  (the order of  $a$ ) does not divide  $qs$ , but  $a^{qs} = e$ , which is a contradiction.  $\square$

**12.F.23.** Prove that every finite group  $G$  whose order is greater than 2 has a non-trivial automorphism.

**Solution.** If  $G$  is not commutative and  $a$  is an element that does not lie in the center, then the conjugation  $x \mapsto axa^{-1}$  defines a non-trivial automorphism. For a cyclic group of order  $m$ , we have, for any  $n$  coprime to  $m$ , the automorphism  $x \mapsto x^n$ . If  $G$  is commutative, then it is a product of cyclic groups (see 10.1.8). If the order of at least one of the factors is greater than 2, then we can use the above automorphism for cyclic groups. If the order of each factor is 2, then permuting any pair of factors is a non-trivial automorphism.  $\square$

**12.F.24.** Consider the group  $(\mathbb{Q}, +)$  of the rational numbers with addition and the group  $(\mathbb{Q}^+, \cdot)$  of the positive rational numbers with multiplication. Find all homomorphisms  $(\mathbb{Q}, +) \rightarrow (\mathbb{Q}^+, \cdot)$ .

**Solution.** There is only one homomorphism, the trivial one. For the sake of contradiction, assume that there exists a non-trivial homomorphism  $\varphi$ , i. e.,  $\varphi(a) = b \neq 1$  for some  $a, b \in \mathbb{Q}$ . Then, for all  $n \in \mathbb{N}$ , we have  $b = \varphi(a) = \varphi(n \frac{a}{n}) = \varphi(\frac{a}{n})^n$ . This is a contradiction, since only some  $n$ -th roots of  $b$  are rational (cf. ??).  $\square$

**12.F.25.** Let  $G$  be the group of matrices of the form  $\begin{pmatrix} a & 0 \\ b & a^{-1} \end{pmatrix}$ , where  $a, b \in \mathbb{R}$  and  $a > 0$ , and let  $N$  be the set of matrices of the form  $\begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}$ , where  $b \in \mathbb{R}$ . Show that  $N$  is a normal subgroup of  $G$  and prove that  $G/N$  is isomorphic to  $\mathbb{R}$ .

**Solution.** The key to the proof is the formula for multiplication in  $G$ :

$$\begin{pmatrix} a & 0 \\ b & a^{-1} \end{pmatrix} \begin{pmatrix} a_1 & 0 \\ b_1 & a_1^{-1} \end{pmatrix} = \begin{pmatrix} aa_1 & 0 \\ ba_1 + a^{-1}b_1 & a^{-1}a_1^{-1} \end{pmatrix}.$$

Hence we can see that the mapping  $\begin{pmatrix} a & 0 \\ b & a^{-1} \end{pmatrix} \mapsto a$  is a homomorphism with kernel  $N$ . Thus,  $N$  is a normal subgroup of  $G$ . Moreover,  $G/N$  is isomorphic to the multiplicative group  $\mathbb{R}^+$ , which is isomorphic to the additive group  $\mathbb{R}$ .  $\square$

**12.F.26.** Let  $G$  be the group with operation of matrix multiplication of  $2 \times 2$  matrices  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$  where  $ad - bc \neq 0$  and  $a, b, c, d$  are integers from  $\mathbb{Z}_3$ . Show that

- i)  $|G| = 48$ .
- ii) if  $ad - bc = 1$  then  $|G| = 24$ .

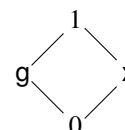
**Solution.**

$(\mathcal{I}_1, \mathcal{I}_2)$	$(0, 0)$	$(0, \times)$	$(\times, 0)$	$(\times, \times)$	$(0, 1)$	$(1, 0)$	$(\times, 1)$	$(1, \times)$	$(1, 1)$
$\mathcal{I}$	0	0	0	$\times$	g	g	1	1	1

We create the interpretation of negation by taking the negation interpretations in  $\mathcal{I}_1$  and  $\mathcal{I}_2$  (recall that if  $|A| = \times$ , then  $|\neg A| = \times$  in Lukasiewicz's three-valued logic). This results in the table:

$A$	$\neg A$
0	1
$\times$	$\times$
g	g
1	0

Furthermore, on the set of truth values, we introduce a (partial) ordering:



and we define the interpretation of propositional connectives using the rules:

$$|A \rightarrow B| = \sup\{|\neg A|, |B|\}, |A \vee B| = \sup\{|A|, |B|\}, |A \wedge B| = \inf\{|A|, |B|\}$$

The semantics is thus a straightforward generalization of the semantics of Lukasiewicz's fuzzy propositional calculus.

Within this semantics, we define a tautology as a statement that, for any interpretation of the statements it is composed of, has an interpretation of 1 or g.

Furthermore, we say that statement  $A$  follows from the set of statements  $\{A_1, \dots, A_n\}$  if, for any interpretation of atomic statements such that statements  $A_1, \dots, A_n$  have an interpretation of 1 or g, statement  $A$  has an interpretation of 1 or g.

If statement  $A$  has an interpretation of  $\times$ , then statements  $A \vee \neg A$  and  $\neg(A \wedge \neg A)$  also have an interpretation of  $\times$  and are not tautologies. In the considered paraconsistent logic, the Law of Excluded Middle does not hold.

If statement  $A$  has an interpretation of g, and statement  $B$  has an interpretation of 1, then statement  $A \rightarrow B$  has an interpretation of 1, and statement  $A \rightarrow \neg B$  has an interpretation of g. This means that in paraconsistent logic, "local inconsistency" can occur (a case where a statement and its negation both follow from another statement). If statement  $A$  has an interpretation of g, then statement  $A \wedge \neg A$  also has an interpretation of g, and some statement  $B$  with an interpretation of 0 or  $\times$  does not follow from it. This means that the Law of Non-Contradiction does not hold in paraconsistent logic, and "global inconsistency" (a case where every statement follows from some set of statements) does not occur.

A propositional logic with "local inconsistencies" but without "global inconsistencies" is interesting because people sometimes hold beliefs that are factually contradictory and yet manage to work with them relatively successfully.

- i) For the first row  $(a, b)$  of a  $M \in G$  values  $a, b$  could be arbitrary except for  $a = 0, b = 0$  in  $\mathbb{Z}_3$ . Hence, there are  $3 \times 3 - 1 = 8$  possibilities to fill the first row. The second row should be not a multiple of the first row. The multiplication factor could be 0, 1 or 2. Hence, for the second row there are  $3 \times 3 - 3 = 6$  possibilities to fill the second row once the first row is filled. Thus,  $|G| = 8 \times 6 = 48$ .
- ii) The set of matrices  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ , such that  $ad - bc = 1$  forms a normal subgroup in  $G$ . Note that for  $\forall g \in G$  values of  $\det g \in \mathbb{Z}_3$  can be either 1 or 2. E.g.  $\det g = 2$  for  $g = \begin{pmatrix} 2 & 2 \\ 1 & 2 \end{pmatrix}$  and  $\det h = 1$  for  $h = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ . Hence,  $\det g' = 2 \quad \forall g' \in Hg$  and, conversely, if  $\det g' = 2$  then  $g' \in Hg$ . Multiplication by  $g$  bijectively maps  $H$  on  $Hg$ . Hence, index  $H \in G$  equals 2 and, therefore,  $|H| = \frac{|G|}{2} = 24$ .

□

**12.F.27.** If  $N$  is a normal subgroup of  $G$  and  $H$  is any subgroup in  $G$ , then  $NH$  is also a subgroup in  $G$ .

**Solution.** Let  $n_1h_1$  and  $n_2h_2$  are two elements in  $NH$  Then

- i)  $n_1h_1n_2h_2 = n_1h_1n_2(h_1)^{-1}h_1h_2 = n_1n_2'h_1h_2 = n'h' \in NH$ .
- ii)  $(nh)^{-1} = h^{-1}n^{-1} = h^{-1}n^{-1}h^{-1}h = n'h$ .

□

**12.F.28.** Suppose that  $N$  and  $M$  are two normal subgroups of  $G$  and  $N \cap M = \{e\}$ . Show that  $nm = mn$  for any  $n \in N$  and  $m \in M$ .

**Solution.** On the one hand,  $mnm^{-1}n^{-1} = m(nm^{-1}n^{-1}) = mm' \in M$ . On the other hand,  $mnm^{-1}n^{-1} = (mnm^{-1})n^{-1} = n'n^{-1} \in N$ . Hence  $mnm^{-1}n^{-1} \in M \cap N$ . So,  $mnm^{-1}n^{-1} = e$ . Thus,  $mn = nm$ . □

**12.F.29.** Show that the intersection of two normal subgroups of  $G$  is also a normal subgroup of  $G$ .


**Solution.** It is clear that the intersection of two subgroups of  $G$  is also a subgroup of  $G$ . Let  $N$  and  $M$  be normal subgroups of  $G$ . Then for any  $g \in G$  and any  $n \in N$  and any  $m \in M$  we have  $gnm^{-1} \in N$  and  $gmg^{-1} \in M$ . Let  $w \in N \cap M$ . Since  $w \in N$  we have  $gwg^{-1} \in N$  and since  $w \in M$  we have  $gwg^{-1} \in M$ . Hence  $gwg^{-1} \in N \cap M$  for any  $g \in G$  i.e.  $N \cap M$  is a normal subgroup of  $G$ . □

**12.F.30.** Prove that if  $G$  is some group and  $H \subset G$  its subgroup of index 2 then  $H$  is a normal subgroup of  $G$ .

**Solution.** Let  $H$  and  $aH$  be the left cosets of  $H$  in  $G$  and  $H$  and  $Hb$  be right cosets of  $H$  in  $G$ . Since there are only two cosets  $aH = G \setminus H$  and  $Hb = G \setminus H$ , then  $aH = Hb$ . In order to show that  $H$  is normal in  $G$ , we need  $xH = Hx$  for any  $x \in G$ . Let  $x \in G$ . If  $x \in H$  then, obviously,

### 3. Polynomial rings

The operations of addition and multiplication are fundamental in the case of scalars as well as vectors. There are other similar structures. Besides the integers  $\mathbb{Z}$ , rational numbers  $\mathbb{Q}$  and complex numbers  $\mathbb{C}$ , there are polynomials over similar scalars  $\mathbb{K}$  to be considered.



Among others, the abstract algebraic theory can be in many aspects viewed as a straightforward generalization of divisibility properties of integers.

**12.3.1. Rings and fields.** Recall that the integers and all other scalars  $\mathbb{K}$  have the following properties:

#### COMMUTATIVE RINGS AND INTEGRAL DOMAINS

**Definition.** Let  $(M, +, \cdot)$  be an algebraic structure with two binary operations  $+$  and  $\cdot$ . It is a *commutative ring*, if it satisfies

- $(a + b) + c = a + (b + c)$  for all  $a, b, c \in M$ ;
- $a + b = b + a$  for all  $a, b \in M$ ;
- there is an element  $0$  such that for all  $a \in M, 0 + a = a$ ;
- for each  $a \in M$ , there is the unique element  $-a \in M$  such that  $a + (-a) = 0$ .
- $(a \cdot b) \cdot c = a \cdot (b \cdot c)$  for all  $a, b, c \in M$ ;
- $a \cdot b = b \cdot a$  for all  $a, b \in M$ ;
- there is an element  $1$  such that for all  $a \in M, 1 \cdot a = a$ ;
- $a \cdot (b + c) = a \cdot b + a \cdot c$  for all  $a, b, c \in M$ .

If the ring is such that  $c \cdot d = 0$  implies either  $c$  or  $d$  is zero, then it is called an *integral domain*.

The first four properties define the algebraic structure of a *commutative group*  $(M, +)$ . Groups are considered in more detail in the next part of this chapter. The last property in the list of ring axioms is called *distributivity* of multiplication over addition. There are similar axioms for Boolean algebras where each of the operations is distributive over the other.

If the operation " $\cdot$ " is commutative for all elements, then the ring is called a commutative ring. Otherwise, the ring is called a non-commutative ring. In the sequel, rings are commutative unless otherwise stated. Traditionally, the operation " $+$ " is called addition, and the operation " $\cdot$ " multiplication, even if they are not the standard operations on one of the known rings of numbers.

In the literature, there are structures without the assumption of having the identity for multiplication. These are not discussed here, so it is always assumed that a ring has a multiplicative identity denoted by  $1$ . The identity for addition is denoted by  $0$ .

#### FIELDS

A non-trivial ring where all non-zero elements are invertible with respect to multiplication is called a *division ring*. If the multiplication is commutative, it is called a *field*.

$xH = H = Hx$  i.e.  $xH = Hx$ . If  $x \in G \setminus H = aH = Hb$  then there exist  $h_1, h_2 \in H$  such that  $x = ah_1 = h_2b$ . Then  $xH = (ah_1)H = aH = Hb = H(h_2b) = Hx$ , i.e.  $xH = Hx$  for any  $x \in G$ , so  $G^{-1}HG = H$ .  $\square$

**12.F.31.** Show that the intersection of two normal subgroups in  $G$  is a normal subgroup of  $G$ .

**Solution.** It is clear that the intersection of two subgroups of  $G$  is a subgroup of  $G$ : Let  $N$  and  $M$  be normal subgroups of  $G$ : Then for any  $g \in G$  and any  $n \in N$  and any  $m \in M$  we have  $gng^{-1} \in N$  and  $gmg^{-1} \in M$ . Let  $w \in N \cap M$ . Since  $w \in N$  we have  $gwg^{-1} \in N$  and since  $w \in M$  we have  $gwg^{-1} \in M$ . Hence  $gwg^{-1} \in N \cap M$  for any  $g \in G$  i.e.  $N \cap M$  is a normal subgroup in  $G$ .  $\square$

**12.F.32.** If  $N$  and  $M$  are normal subgroups of  $G$  show that  $NM$  is also a normal subgroup of  $G$ .

**Solution.** Let  $nm \in NM$ , then

$$gnmg^{-1} = gng^{-1}gmg^{-1} \in NM,$$

since  $gmg^{-1} \in M$  and  $gng^{-1} \in N$ .  $\square$

**12.F.33.** If  $N$  is a normal subgroup in the finite group such that  $\gcd(|G : N|, |N|) = 1$ . Show that  $N$  must contain any element  $x \in G$  satisfying  $x^{|N|} = e$ .

**Solution.** Let  $x$  be an element in  $G$  such that  $x^{|N|} = e$ . Since  $\gcd(|G : N|, |N|) = 1$  there exist  $m, n \in \mathbb{Z}$  such that

$$m|G : N| + n|N| = 1.$$

Then  $x = x^{m|G:N| + n|N|} = x^{m|G:N|} x^{n|N|}$ . Since  $x^{|N|} = e$  we have  $x = x^{m|G:N|}$ . Consider element  $xN \in G/N$ . Since  $xN^k = x^kN$  for any  $k \in \mathbb{Z}$ , then  $N = (xN)^{|G/N|} = x^{|G/N|}N$ , which is the identity element in  $G/N$ . This means  $x^{|G/N|} \in N$  and so  $x = (x^{|G/N|})^m \in N$ .  $\square$

**12.F.34.** If  $H$  is a subgroup of  $G$  such that the product of any two right cosets of  $H$  in  $G$  is again a right coset of  $H$  in  $G$  show that  $H$  is normal in  $G$ .

**Solution.** Let  $Ha$  and  $Hb$  be two right cosets of  $H \in G$ . By assumption  $HaHb$  is a right coset of  $H$  in  $G$ : The set  $HaHb$  contains the element  $eaeb = ab$ . Since there is only one right coset of  $H$  in  $G$  containing  $ab$ ; which is  $Hab$ ; we have  $HaHb = Hab$  for all  $a, b \in G$ . Hence  $HaH = Ha$  for all  $a \in G$ . i.e.  $h_1ah_2$  is equal to  $h_3a$  for some  $h_3 \in H$ . But  $h_1ah_2 = h_3a$  implies  $ah_2a^{-1} = h_1^{-1}h_3 = h' \in H$  for any  $h_2 \in H$  and  $a \in G$ . So,  $H$  is normal in  $G$ .  $\square$

**12.F.35.** Let  $Z(G)$  be centraliser of group  $G$  and  $xy = z \in Z(G)$  for some  $x, y \in G$ . Show that  $x$  and  $y$  commute.

**Solution.** As  $zx = xz$ , we have  $xyx = zx = xz = xxy$ . Multiplying by  $x^{-1}$  on the left, we obtain  $yx = xy$ .  $\square$

Typical examples of fields are the rational numbers  $\mathbb{Q}$ , the real numbers  $\mathbb{R}$ , and the complex numbers  $\mathbb{C}$ . Furthermore, every remainder class set  $\mathbb{Z}_p$  is a commutative ring, while only  $\mathbb{Z}_p$  for prime  $p$  are also fields. We shall come back to fields in the next part of this chapter, too.

Recall the useful example of a non-commutative ring, the set  $\text{Mat}_k(\mathbb{K})$  of all  $k$ -by- $k$  matrices over a ring  $\mathbb{K}$ ,  $k \geq 2$ . As can be checked for  $\mathbb{K} = \mathbb{Z}_2$  and  $k = 2$ , these rings are never an integral domain (see 2.1.5 on page 87 for the full argument).

As an example of a division ring which is not a field, consider the ring of quaternions  $\mathbb{H}$ . This is constructed as an extension of the complex numbers, by adding another imaginary unit  $j$ , i.e.  $\mathbb{H} = \mathbb{C} \oplus j\mathbb{C} \simeq \mathbb{R}^4$ , just as the complex numbers are obtained from the reals. Another “new” element  $i \cdot j$  is usually denoted  $k$ . It follows from the construction that  $ij = -ji$ . This structure is a division ring. Think out the details as a not completely easy exercise!

**12.3.2. Elementary properties of rings.** The following



lemma collects properties which all seem to be obvious for rings of scalars. But the properties need proof to build an abstract

theory:

**Lemma.** In every commutative ring  $\mathbb{K}$ , the following holds

- (1)  $0 \cdot c = c \cdot 0 = 0$  for all  $c \in \mathbb{K}$ ,
- (2)  $-c = (-1) \cdot c = c \cdot (-1)$  for all  $c \in \mathbb{K}$ ,
- (3)  $-(c \cdot d) = (-c) \cdot d = c \cdot (-d)$  for all  $c, d \in \mathbb{K}$ ,
- (4)  $a \cdot (b - c) = a \cdot b - a \cdot c$ ,
- (5) the entire ring  $\mathbb{K}$  collapses to the trivial set  $\{0\} = \{1\}$  if and only if  $0 = 1$ .

**PROOF.** All of the propositions are direct consequences of the definition axioms. In the first case, for any  $c, a$

$$c \cdot a = c \cdot (a + 0) = c \cdot a + c \cdot 0,$$

and since  $0$  is the only element that is neutral with respect to addition,  $c \cdot 0 = 0$ .

In the second case, it suffices to compute

$$0 = c \cdot 0 = c \cdot (1 + (-1)) = c + c \cdot (-1).$$

This means that  $c \cdot (-1)$  is the inverse of  $c$ , as desired.

The following two propositions are direct consequences of the second proposition and the axioms. If the ring contains only one element, then  $0 = 1$ . On the other hand, if  $1 = 0$ , then for any  $c \in \mathbb{K}$ , necessarily  $c = 1 \cdot c = 0 \cdot c = 0$ .  $\square$

**12.3.3. Polynomials over rings.** The definition of a commu-



tative ring uses precisely the properties that are expected for multiplication and addition. The concept of polynomials can now be extended. A *polynomial* is any expression that can be built from (known) constant elements of  $\mathbb{K}$  and an (unknown) variable using finite number of additions and multiplications. Formally, polynomials are defined as follows:<sup>4</sup>

<sup>4</sup>It is not by accident that the symbol  $\mathbb{K}$  is used for the ring – you can imagine e.g. any of the number rings behind that.

**12.F.36.** Let  $G$  be a group of  $2 \times 2$  matrices  $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  such that  $ad - bc \neq 0, a, b, c, d \in \mathbb{R}$ . Let subgroup  $H = \left\{ \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \mid x \in \mathbb{Z} \right\} \in G$  and  $g = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ . Show that  $H \neq gHg^{-1} \subset H$ .

**Solution.** Obviously, matrix  $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in H$ . However,

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2x & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2x & 1 \end{pmatrix}$$

Thus,  $g^{-1}Hg = \left\{ \begin{pmatrix} 1 & 0 \\ 2x & 1 \end{pmatrix} \mid x \in \mathbb{Z} \right\}$  does not contain  $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in H$ .  $\square$

**12.F.37.** Let  $G$  be a group of an even order. Prove that it contains a subgroup of order 2.

**Solution.** Consider the set of pairs  $\Omega = \{g, g^{-1}\}$ , where  $g \neq g^{-1}$ . This engages an even number of elements of  $G$ . Those elements  $g \in G$  that are not in  $\Omega$  are the ones satisfying  $g = g^{-1}$  i.e.,  $g^2 = e$ . Therefore if we count  $G \bmod 2$ , we can ignore the pairs  $\{g, g^{-1}\} \in \Omega$  and obtain  $|G| \equiv |\{g \in G \mid g^2 = e\}| \pmod{2}$ . One solution to  $g^2 = e$  is  $e$  itself. If it were the only solution, then  $|G| \equiv 1 \pmod{2}$ , which is false. Therefore some  $g \neq e$  satisfies  $g^2 = e$ , which provides an element and also a subgroup in  $G$  of order 2.  $\square$

**12.F.38.** Let  $G$  be finite abelian group, which order  $n$  is divisible by prime factor  $p$ . Show that  $G$  contains an element of order  $p$  and, hence a cyclic subgroup of that order. This claim is known as abelian case of Cauchy theorem.

**Solution.** We use induction on  $n$ . Case  $n = p$  is trivial. Let  $n > p$  and  $p|n$  and the proposition is true for all abelian groups of orders divisible by  $p$  and less than  $n$ . Assume no element of  $G$  has order  $p$ . Then no element has order divisible by  $p$ , because if  $g \in G$  has order  $r$  and  $p|r$  then  $g^{\frac{r}{p}}$  would have order  $p$ . Let  $G = \{g_1, g_2, \dots, g_n\}$  and let  $g_i$  have order  $m_i$  so  $m_i$  is not divisible by  $p$ . Set  $m$  to be the least common multiple of all  $m_i, i = 1, 2, \dots, n$ , so  $m$  is not divisible by  $p$  and  $g^{m_i} = e$  for all  $i$ . Because  $G$  is abelian, the function  $f : (\mathbb{Z}/m)^n \rightarrow G$  given by

$$f(a_1, \dots, a_n) = g_1^{a_1} \dots g_n^{a_n}.$$

is a homomorphism  $f(a_1, \dots, a_n)f(b_1, \dots, b_n) = f(a_1 + b_1, \dots, a_n + b_n)$ . as, by commutativity of  $G$ ,

$$g_1^{a_1} \dots g_n^{a_n} g_1^{b_1} \dots g_n^{b_n} = g_1^{a_1+b_1} \dots g_n^{a_n+b_n}$$

's. This homomorphism is surjective (each element of  $g_i \in G$  is  $F(a_1, \dots, a_n)$  for  $a_i = 1$  and other  $a_j = 0, j \neq i$ .  $g_i$ , and if  $a_i = 1$  and other  $a_j$ 's are 0 then  $f(a_1, \dots, a_n) = g_i$ ) and the elements where  $f$  takes particular value  $g_i$  is a coset of  $\ker f$ , so  $|G| = \text{number of cosets of } \ker f$  which is a factor of  $|\mathbb{Z}/m|^n$ , which is a factor of  $m^n$ . Since  $p||G|$  and  $m^n$  is not divisible by  $p$ , this is a contradiction.  $\square$

POLYNOMIALS

**Definition.** Let  $\mathbb{K}$  be a commutative ring. A polynomial over  $\mathbb{K}$  is a finite expression

$$f(x) = \sum_{i=0}^k a_i x^i,$$

where  $a_i \in \mathbb{K}, i = 0, 1, \dots, k$  are the *coefficients of the polynomial*. If  $a_k \neq 0$ , then by definition,  $f(x)$  has *degree*  $k$ , written  $\deg f = k$ . The zero polynomial is not assigned a degree. Polynomials of degree zero (called *constant polynomials*) are exactly the non-zero elements of  $\mathbb{K}$ .

Polynomials  $f(x)$  and  $g(x)$  are equal if they have the same coefficients. The set of all polynomials over a ring  $\mathbb{K}$  is denoted  $\mathbb{K}[x]$ .

Every polynomial defines a mapping  $f : \mathbb{K} \rightarrow \mathbb{K}$  by substituting the argument  $c$  for the variable  $x$  and evaluating the resulting expression, i.e.

$$f(c) = a_0 + a_1c + \dots + a_kc^k.$$

Note that constant polynomials define constant mappings in this manner.

A *root of a polynomial*  $f(x)$  is such an element  $c \in \mathbb{K}$  for which  $f(c) = 0 \in \mathbb{K}$ .

It may happen that different polynomials define the same mapping. For instance, the polynomial  $x^2 + x \in \mathbb{Z}_2[x]$  defines the mapping which is constantly equal to zero. More generally, for every finite ring  $\mathbb{K} = \{a_0, a_1, \dots, a_k\}$ , the polynomial  $f(x) = (x - a_0)(x - a_1) \dots (x - a_k)$  defines the constant-zero mapping.

Polynomials  $f(x) = \sum_i a_i x^i$  and  $g(x) = \sum_i b_i x^i$  can be added and multiplied in a natural way (just think to introduce again the structure of a ring and invoke the expected distributivity of multiplication over addition):

$$\begin{aligned} (f + g)(x) &= (a_0 + b_0) + (a_1 + b_1)x + \dots + (a_k + b_k)x^k, \\ (f \cdot g)(x) &= (a_0b_0) + (a_0b_1 + a_1b_0)x + \dots \\ &\quad + (a_0b_r + a_1b_{r-1} + a_r b_0)x^r + \dots + a_k b_\ell x^{k+\ell}, \end{aligned}$$

where  $k \geq \ell$  are the degrees of  $f$  and  $g$ , respectively. Zero coefficients are assumed everywhere where there is no coefficient in the original expression.<sup>5</sup>

This definition corresponds to the addition and multiplication of the function values of  $f, g : \mathbb{K} \rightarrow \mathbb{K}$ , by the properties of "coefficients" in the original ring  $\mathbb{K}$ .

It follows directly from the definition that the set of polynomials  $\mathbb{K}[x]$  over a commutative ring  $\mathbb{K}$  is again a commutative ring, where the multiplicative identity is the element  $1 \in \mathbb{K}$ , perceived as a polynomial of degree zero. The additive identity is the zero polynomial. You should check all the axioms carefully!

<sup>5</sup>To avoid this formal hassle, a polynomial can be defined as an infinite expression (like a formal power series over the ring in question) with the condition that only finitely many coefficients are non-zero. Concerning the order, zero polynomial is also attributed to have order  $-\infty$ .



**12.F.39.** Let  $G$  be finite -abelian group, which order  $n$  is divisible by prime factor  $p$ . Show that  $G$  contains an element of order  $p$  and, hence a cyclic subgroup of that order. This is non-abelian case of Cauchy theorem.

**Solution.** If a proper subgroup  $H$  of  $G$  has order divisible by  $p$ , then by induction there is an element of order  $p$  in  $H$ , which gives us an element of order  $p$  in  $G$ . Thus we may assume no proper subgroup of  $G$  has order divisible by  $p$ . For any proper subgroup  $H \subset G$  we have  $|G| = |H|[G : H]$  and  $|H|$  is not divisible by  $p$ , so  $p \mid [G : H]$  for every proper subgroup  $H$ . Denote conjugacy class of an element  $a \in G$  by  $C(a) = \{b \in G \mid b = g^{-1}ag \text{ for some } g \in G\}$ . Let the conjugacy classes in  $G$  with size greater than 1 be represented by  $g_1, g_2, \dots, g_k$ . The conjugacy classes of size 1 are the elements of the center  $Z(G)$ . Since the conjugacy classes are a partition of  $G$ , counting  $|G|$  by counting conjugacy classes implies

$$|G| = |Z(G)| + \sum_{i=1}^k |C(g_i)| = |Z(G)| + \sum_{i=1}^k [G : Z(g_i)],$$

where  $Z(g_i)$  is the centraliser of  $g_i$ . Every  $[G : Z(g_i)] > 1$  since each  $C(g_i)$  is greater than 1. Since  $\forall j, |G| = |Z(g_i)[G : Z(g_i)]$ , it yields  $p \mid [G : Z(g_i)]$ . Thus,  $|Z(G)|$  is divisible by  $p$  and so is non-trivial. Since proper subgroups of  $G$  do not have order divisible by  $p$ , center  $Z(G)$  has to be all of  $G$ . That means  $G$  is abelian, which is a contradiction.  $\square$

**12.F.40.** Prove that any group of order 15 is cyclic.

**Solution.** Let  $G$  be a group of order 15. By Cauchy's theorem there exists a subgroup  $M$  of order 5 and a subgroup  $N$  of order 3.  $[G : M] = \frac{|G|}{|M|} = 3$ . Both  $M$  and  $N$  are cyclic subgroups. Let  $x$  be a generator of  $M$  and  $y$  - the generator of  $N$ . Let's prove that  $xy$  generates entire  $G$ . (NEED TO COME UP WITH AN ELEGANT PROOF..)

$$(xg)^5 =$$

$\square$

**12.F.41.** Let  $G$  be a group of order 14 which has a normal subgroup  $N$  of order 2. Prove that  $G$  is commutative.

**Solution.** Clearly, the order of the group  $G/N$  is  $|G/N| = \frac{|G|}{|N|} = 7$ . By Lagrange's theorem 12.4.10, the orders of its elements are 1 or 7. Since only the identity has order 1, this means that there is an element of order 7, so that the group  $G/N$  is cyclic. Let  $N = \{e, n\}$ , where  $e$  is the identity of  $G$  and let  $[a]$  be the generator of  $G/N$ . Since  $N$  is normal, we have  $ana^{-1} \in N$ , but  $ana^{-1} = e$  implies  $n = e$ , which means that we must have  $ana^{-1} = n$ , i. e.,  $na = an$ . Since  $[a]$  generates  $G/N$ , we get that each element of  $G/N$  is of the form  $[a]^k, k = 0, \dots, 6$ , i. e.,  $[a^k]$ . Then, each element of  $G$

**Lemma.** A polynomial ring over an integral domain is again an integral domain.

**PROOF.** The task is to show that  $\mathbb{K}[x]$  can contain non-trivial divisors of zero only if they are in  $\mathbb{K}$ . However, this is clear from the expression for polynomial multiplication. If  $f(x)$  and  $g(x)$  are polynomials of degree  $k$  and  $\ell$  as above, then the coefficient at  $x^{k+\ell}$  in the product  $f(x) \cdot g(x)$  is the product  $a_k \cdot b_\ell$ , which is non-zero unless there are zero divisors in  $\mathbb{K}$ .  $\square$

**12.3.4. Multivariate polynomials.** Some objects can be described using polynomials with more variables.

For instance, consider a circle in the plane  $\mathbb{R}^2$  whose center is at  $S = (x_0, y_0)$  and whose radius is  $R$ . This circle can be defined by the equation

$$(x - x_0)^2 + (y - y_0)^2 - R^2 = 0.$$

Rings of polynomials in variables  $x_1, \dots, x_r$  can be defined similarly as in the case of  $\mathbb{K}[x]$ . Instead of the powers  $x^k$  of a single variable  $x$ , consider the *monomials*

$$x_1^{k_1} \dots x_r^{k_r},$$

and their formal linear combinations with coefficients  $a_{k_1 \dots k_r} \in \mathbb{K}$ .

However, it is simpler, both formally and technically, to define them inductively by

$$\mathbb{K}[x_1, \dots, x_r] := (\mathbb{K}[x_1, \dots, x_{r-1}])[x_r].$$

For instance,  $\mathbb{K}[x, y] = \mathbb{K}[x][y]$ . One can consider polynomials in the variable  $y$  over the ring  $\mathbb{K}[x]$ . It can be shown (check this in detail!) that polynomials in variables  $x_1, \dots, x_r$  can be viewed, even with this definition, as expressions created from the variables  $x_1, \dots, x_n$  and the elements of the ring  $\mathbb{K}$  with a finite number of (formal) addition and multiplication in a commutative ring. For example, the elements of  $\mathbb{K}[x, y]$  are of the form

$$\begin{aligned} f &= a_n(x)y^n + a_{n-1}(x)y^{n-1} + \dots + a_0(x) = \\ &= (a_{mn}x^m + \dots + a_{0n})y^n + \dots + (b_{p0}x^p + \dots + b_{00}) = \\ &= c_{00} + c_{10}x + c_{01}y + c_{20}x^2 + c_{11}xy + c_{02}y^2 + \dots \end{aligned}$$

To simplify the notation, we use the multi-index notation (as we did with real polynomials and partial derivatives in infinitesimal analysis).

is of the form  $a^k$  or  $a^k n$ , and since  $a$  and  $n$  commute, we get that actually all elements of  $G$  commute.  $\square$

**12.F.42.** Decide whether the following holds: If the quotient  $G/N$  of a group  $G$  by a normal subgroup  $N$  is commutative, then  $G$  itself is commutative.  $\circ$

**12.F.43.** Prove that any subgroup  $H$  of the symmetric group  $\mathbb{S}_n$  contains either only even permutations or the same number of even and odd permutations.

**Solution.** Consider the homomorphism  $p : H \rightarrow \mathbb{Z}_2$  which maps each permutation to its parity (0 for even and 1 for odd). Then,  $p^{-1}(0) = \text{Ker}(p)$  is a normal subgroup of  $H$ : let  $h \in \text{Ker}(p)$ , then

$$p(ghg^{-1}) = p(g)p(h)p(g^{-1}) = p(g)p(g^{-1}) = p(gg^{-1}) = p(e) = 0,$$

which means that  $ghg^{-1} \in \text{Ker}(p)$ , i. e.,  $\text{Ker}(p)$  is normal. Since  $\mathbb{Z}_2$  has only two elements, it follows that  $H/\text{Ker}(p)$  has either only one coset (i. e., all permutations are even) or two cosets, which must be of equal size (i. e., there are the same number of even and odd permutations).  $\square$

**12.F.44.** Describe the group of symmetries of a regular tetrahedron and find all of its subgroups.

**Solution.** Let us denote the vertices of the tetrahedron as  $a, b, c, d$ . Each symmetry can be described as a permutation of the vertices (to which vertex each one goes). Thus, the group of symmetries of the tetrahedron is isomorphic to a certain subgroup of the symmetric group  $\mathbb{S}_4$ . Given any pair of vertices, there exists a symmetry which swaps this pair and keeps the other two vertices fixed (this is reflection with respect to the plane that is perpendicular to the line segment of the pair and goes through its center). Thus, the wanted subgroup is generated by all transpositions in  $\mathbb{S}_4$ . However, this is the group  $\mathbb{S}_4$  itself.

Thus, let us describe all subgroups of the group  $\mathbb{S}_4$ . This group has 24 elements, which means that the order of any subgroup must be one of 1, 2, 3, 4, 6, 8, 12, 24 (see 12.4.10). Clearly, the only group of order 1 is the trivial subgroup  $\{id\}$ . Similarly, the only group of order 24 is the entire group  $\mathbb{S}_4$ . Now, let us look at the remaining orders of a potential subgroup  $H \subseteq \mathbb{S}_4$ .

(i)  $|H| = 2$ .  $H$  must consist of the identity and another self-inverse element ( $x^2 = id$ ). These are transpositions and double transpositions (compositions of two disjoint transpositions). Geometrically, double transpositions correspond to rotation by  $180^\circ$  around the axis that goes through the centers of opposite edges). Thus, we get nine subgroups:  $\{id, (a, b)\}$ ,  $\{id, (a, c)\}$ ,  $\{id, (a, d)\}$ ,  $\{id, (b, c)\}$ ,  $\{id, (b, d)\}$ ,  $\{id, (c, d)\}$ ,  $\{id, (a, b) \circ (c, d)\}$ ,  $\{id, (a, c) \circ (b, d)\}$ ,  $\{id, (a, d) \circ (b, c)\}$ .

(ii)  $|H| = 3$ . By Lagrange's theorem, such a subgroup must be cyclic, i. e., it must be of the form  $\{id, p, p^2\}$ ,  $p^3 = id$ . Thus, the factorization of  $p$  to independent cycles must contain a cycle of length 3, which means that  $p$  cannot contain anything else. By 12.F.17, there are  $4 \cdot 2$  cycles of length 3, which give rise to the following four

MULTI-INDICES

A *multi-index*  $\alpha$  of length  $r$  is an  $r$ -tuple of non-negative integers  $(\alpha_1, \dots, \alpha_r)$ . The integer  $|\alpha| = \alpha_1 + \dots + \alpha_r$  is called the *size* of the multi-index  $\alpha$ .

Monomials are written shortly as  $x^\alpha$  instead of  $x_1^{\alpha_1} x_2^{\alpha_2} \dots x_r^{\alpha_r}$ . Polynomials in  $r$  variables can be symbolically expressed in a similar way as univariate polynomials:

$$f = \sum_{|\alpha| \leq n} a_\alpha x^\alpha, \quad g = \sum_{|\beta| \leq m} b_\beta x^\beta \in \mathbb{K}[x_1, \dots, x_r].$$

$f$  is said to have total degree  $n$  if at least one coefficient with multi-indices  $\alpha$  of size  $n$  is non-zero, while all the coefficients with multi-indices of larger sizes vanish.

Analogous formulae can be defined for addition and multiplication of multivariate polynomials of degrees  $m$  and  $n$  respectively:

$$f + g = \sum_{|\alpha| \leq \max(m, n)} (a_\alpha + b_\alpha) x^\alpha,$$

$$fg = \sum_{|\gamma|=0}^{m+n} \left( \sum_{\alpha+\beta=\gamma} a_\alpha b_\beta \right) x^\gamma,$$

where the multi-indices are added componentwise, and the formally non-existing coefficients are assumed to be zero. Clearly in the case of univariate polynomials we recover the formula

$$fg(x) = \sum_{i=0}^{k+\ell} \left( \sum_{p+q=i} a_p b_q \right) x^i.$$

**Lemma.** *These formulae describe addition and multiplication in the inductively defined ring of polynomials in  $r$  variables. In particular, in a polynomial ring over integral domain, the total degree of a sum or difference of two polynomials (if defined) is at most the maximum of their total degrees, while the total degree of a product of two polynomials is the sum of their total degrees.*

**PROOF.** The proposition is easily proved by induction on the number of variables. Suppose that the formulae are valid in  $\mathbb{K}[x_1, \dots, x_{r-1}]$ , and calculate the sum



$$f = a_k(x_1, \dots, x_{r-1})x_r^k + \dots + a_0(x_1, \dots, x_{r-1})$$

$$= \left( \sum_{\alpha} a_{k, \alpha} x^\alpha \right) x_r^k + \dots,$$

$$g = b_l(x_1, \dots, x_{r-1})x_r^l + \dots + b_0(x_1, \dots, x_{r-1})$$

$$= \left( \sum_{\beta} b_{l, \beta} x^\beta \right) x_r^l + \dots,$$

$$f + g = (a_0(x_1, \dots, x_{r-1}) + b_0(x_1, \dots, x_{r-1}))$$

$$+ (a_1(x_1, \dots, x_{r-1}) + b_1(x_1, \dots, x_{r-1}))x_r + \dots$$

subgroups:  $\{\text{id}, (a, b, c), (a, c, b)\}$ ,  $\{\text{id}, (a, c, d), (a, d, c)\}$ ,  $\{\text{id}, (a, b, d), (a, d, b)\}$ ,  $\{\text{id}, (b, c, d), (b, d, c)\}$ . The cycles of length 3 correspond to rotation by  $120^\circ$  around the axis that goes through a vertex and the center of the opposite side.

(iii)  $|H| = 4$ . Such a subgroup must be isomorphic to  $\mathbb{Z}_4$  or  $\mathbb{Z}_2 \times \mathbb{Z}_2$ . Considering the factorization to independent cycles, we find out that the only permutation of order 4 is a cycle of length 4. Thus, cyclic subgroups must contain a cycle of length 4, namely exactly two of them, since if  $p$  has order 4, then  $p^{-1} = p^3$  is also of order 4 i. e., a cycle of length 4. Then, the permutation  $p^2$  has order 2, so it must be a double transposition (it is not a single transposition, since  $p^2$  clearly does not have a fixed point). There are six cycles of length 4 (see 12.F.17), and they pair up to the following three subgroups of this type:

$$\begin{aligned} &\{\text{id}, (a, b, c, d), (a, c) \circ (b, d), (a, d, c, b)\}, \\ &\{\text{id}, (a, c, b, d), (a, b) \circ (b, d), (a, d, b, c)\}, \\ &\{\text{id}, (a, b, d, c), (a, d) \circ (bc), (a, c, d, b)\}. \end{aligned}$$

As for subgroups isomorphic to  $\mathbb{Z}_2 \times \mathbb{Z}_2$ , they must contain (besides the identity) only elements of order 2, which are transpositions and double transpositions. By 12.F.43, the subgroup must contain either no or exactly two transpositions. Moreover, it cannot contain two dependent transpositions, since their composition is a cycle of length 3. Thus, the subgroup contains (besides the identity) either two independent transpositions and the double transposition which is their composition (this gives rise to three subgroups), or the three double transpositions. Altogether, we have found:  $\{\text{id}, (a, b), (a, b) \circ (c, d), (c, d)\}$ ,  $\{\text{id}, (a, c), (a, c) \circ (b, d), (b, d)\}$ ,  $\{\text{id}, (a, d), (a, d) \circ (b, c), (b, c)\}$  and  $\{\text{id}, (a, b) \circ (c, d), (a, c) \circ (b, d), (a, d) \circ (b, c)\}$ .

(iv)  $|H| = 6$ . By 12.F.11, this subgroup is isomorphic to  $\mathbb{S}_3$  (it cannot be isomorphic to  $\mathbb{Z}_6$  since there is no element of order 6 in  $\mathbb{S}_4$ ), so it contains (besides the identity) two elements  $x, x^{-1}$  of order 3 and three elements of order 2. Thus,  $x$  and  $x^{-1}$  are cycles of length 3 which fix the same vertex (say  $a$ ). What are the other three elements? There cannot be a double transposition, since its composition with  $x$  yields another cycle of length 3. There cannot be a transposition which does not fix  $a$  since its composition with  $x$  yields a cycle of length 4. Thus, the only possibility is that there are the three transpositions which also fix  $a$ . Since there are four possibilities which vertex is the fixed one, we obtain four subgroups of order 6.

(v)  $|H| = 8$ . The group cannot be a subgroup of the group  $\mathbb{A}_4$  of even permutations (since there are 12 of them, and 8 does not divide 12). Thus, by 12.F.43,  $H$  must contain four even and four odd permutations. The even permutations must form a subgroup of  $\mathbb{A}_4$ , and we could see in (iii) that the only such 4-element subgroup is  $\{\text{id}, (a, b) \circ (c, d), (a, c) \circ (b, d), (a, d) \circ (b, c)\}$ , which is normal. Considering any odd permutation and the coset (with respect to the above normal subgroup) which contains it, we can see that the coset together with the above 4 elements form a subgroup of  $\mathbb{S}_4$ . We thus get three subgroups of  $\mathbb{S}_4$ . It is not hard to realize that

$$\begin{aligned} &= \left( \sum_{\gamma} (a_{k,\gamma} + b_{k,\gamma})(x_1, \dots, x_{r-1})^{\gamma} \right) x_r^k + \dots \\ &+ \left( \sum_{\gamma} (a_{0,\gamma} + b_{0,\gamma})(x_1, \dots, x_{r-1})^{\gamma} \right) \\ &= \sum_{(\gamma,j)} (a_{j,\gamma} + b_{j,\gamma})(x_1, \dots, x_{r-1})^{\gamma} x_r^j. \end{aligned}$$

The proof for multiplication is similar (do it yourselves!).  $\square$

The definition or the above formulae for polynomials over general commutative rings yield the following corollary:

**Corollary.** *If a ring  $\mathbb{K}$  is an integral domain, then the ring  $\mathbb{K}[x_1, \dots, x_r]$  is also an integral domain.*

**PROOF.** Proceed by induction on the number  $r$  of variables.<sup>6</sup> We know the result for the univariate polynomials already. In particular, the product of non-zero polynomials is again non-zero.

If the proposition holds for  $r-1$  variables, then the result follows from the inductive definition of the polynomials.  $\square$

Let us notice that each polynomial  $f \in \mathbb{K}[x_1, \dots, x_n]$  also represents a mapping  $\mathbb{K} \times \dots \times \mathbb{K} \rightarrow \mathbb{K}$  multilinear and symmetric in its  $n$  arguments.

**12.3.5. Divisibility and irreducibility.** The next goal is to understand how polynomials over a general integral domain can be expressed as products of simpler polynomials. In the case of univariate polynomials, this means finding the roots of a polynomial. Since multivariate polynomials can be defined inductively, it suffices to consider univariate polynomials over a general integral domain. This leads to a generalization of the concept of divisibility which forms the basis of elementary number theory in chapter eleven.

Consider an integral domain  $\mathbb{K}$  (for instance, the integers  $\mathbb{Z}$  or the ring  $\mathbb{Z}_p$  for prime  $p$ ).



DIVISIBILITY IN RINGS

For  $a, c \in \mathbb{K}$  we say that  $a$  divides  $c$  in  $\mathbb{K}$  if and only if there is some  $b \in \mathbb{K}$  such that  $a \cdot b = c$ . This is written  $a|c$ .

The divisors of 1 (the multiplicative identity), i.e. the invertible elements of  $\mathbb{K}$ , are called *units*.

The units of a commutative ring always form a commutative group in the sense used for the properties of addition in the definition of rings. This is called the group of units in  $\mathbb{K}$ . The group of units in  $\mathbb{Z}$  is  $\{-1, 1\}$ , while all nonzero elements in a field are units there.

In an integral domain, the divisors are determined uniquely. If  $b = a \cdot c$  and  $b \neq 0$ , then  $c$  is determined by the choice of  $a$  and  $b$ , since if  $b = ac = ac'$ , then  $0 = a \cdot (c - c')$

<sup>6</sup>Alternatively, proceed directly using multi-index formulae for the product, provided an appropriate ordering on monomials is defined, see the last part of this chapter.

each of them is isomorphic to the group of symmetries of a square (the so-called dihedral group  $D_4$ ). From the geometrical point of view, we can describe it as follows: Consider the orthogonal projection of the tetrahedron onto the plane that is perpendicular to the line that goes through the centers of opposite edges. The boundary of this projection is a square. Out of all the symmetries of the tetrahedron, we take only those which induce a symmetry of this square (for instance, it will not be a symmetry which only swaps adjacent vertices of the resulting square). Since there are three pairs of opposite edges in the tetrahedron, we get three 8-element subgroups, isomorphic to the dihedral group  $D_4$ . (vi)  $|H| = 12$ . By 12.F.43, such a subgroup contains either only even permutations, or six even and six odd permutations, and the six even permutations must form a subgroup of  $S_4$ . However, we could see in (iv) that there is no 6-element subgroup of  $S_4$  consisting only of even permutations. Thus, the only possibility is the alternating group  $A_4$  of all even permutations in  $S_4$ . From the geometric point of view, these are the so-called direct symmetries, which are realized by rotations (not reflections), and thus can be performed in the space.  $\square$

**Remark.** In general, the group of symmetries of a solid with  $n$  vertices is a subgroup of the symmetric group  $S_n$ .

**12.F.45.** Which subgroups of the group  $S_4$  are normal?

**Solution.** By definition, a subgroup  $H \subseteq S_4$  is normal iff it is closed under conjugation, i. e.,  $ghg^{-1} \in H$  for any  $g \in S_4$ ,  $h \in H$ . Since conjugation in symmetric groups only renames the permuted items but preserves the permutation structure (i. e. the cycle lengths in the factorization to independent cycles), we can see that  $H$  is normal if and only if it contains either no or all permutations of each type. Examining all the subgroups, which we found in the previous exercise, we find that the normal ones are the trivial group  $\{id\}$ , the so-called Klein group (consisting of the identity and the three double transpositions, which we already met in 12.F.8), the alternating group  $A_4$  of all even permutations, and the entire group  $S_4$ .  $\square$

**12.F.46.** Find the group of symmetries of a cube (describe all symmetries). Is this group commutative?

**Solution.** The group has 48 elements; 24 of them are generated by rotations (these are the so-called direct symmetries), the other 24 are the composition of a direct symmetry and a reflection. The group is not commutative (consider the composition of a reflection with respect to the plane containing the centers of four parallel edges and a rotation by  $90^\circ$  around the axis that lies in the plane and goes through the centers of two opposite sides. The group is isomorphic to  $S_4 \times Z_2$ .  $\square$

**12.F.47.** In the group of symmetries of a cube, find the subgroup generated by a reflection with respect to the plane containing the centers of four parallel edges and the rotation by  $180^\circ$  around the axis that lies in the plane and goes through the centers of two opposite sides. Is this subgroup normal?

$\circ$

and being in an integral domain,  $a \neq 0$  implies  $c = c'$ . The reader should check that the claim is wrong e.g. in  $Z_4$ .

Just as for integers, the following propositions are direct corollaries of the definitions (check the details yourself!):

**Lemma.** Let  $a, b, c \in \mathbb{K}$ . Then,

- (1) if  $a|b$  and  $b|c$ , then  $a|c$ ;
- (2) if  $a|b$  and  $a|c$ , then  $a|(\alpha b + \beta c)$  for all  $\alpha, \beta \in \mathbb{K}$ ;
- (3)  $a|0$  (since  $a \cdot 0 = 0$ , in particular  $0|0$ );
- (4)  $a \in \mathbb{K}$  is divisible by each unit  $e \in \mathbb{K}$  and its  $a$ -multiple  $a \cdot e$  (this follows from the existence of  $e^{-1}$ ).

#### UNIQUE FACTORIZATION DOMAIN

An element  $a \in \mathbb{K}$  is said to be *irreducible* if and only if it is divisible only by units  $e \in \mathbb{K}$  and their  $a$ -multiples.

A ring  $\mathbb{K}$  is called a *unique factorization domain* if and only if the following conditions hold:

- For every non-zero element  $a \in \mathbb{K}$ , there are irreducible elements  $a_1, \dots, a_r \in \mathbb{K}$  such that  $a = a_1 \cdot a_2 \cdot \dots \cdot a_r$ . This is called *factorization* of  $a$ .
- If there are two factorizations of  $a$  into irreducible non-unit elements  $a = a_1 a_2 \dots a_r = b_1 b_2 \dots b_s$ , then  $r = s$  and, up to a permutation of the order of the factors  $b_j$ ,  $a_j = e_j b_j$  for suitable units  $e_j, j = 1, \dots, r$ .

$Z$  is a unique factorization domain. So is every field, since every non-zero element is a unit.

There are examples of integral domains without the unique factorization property. The construction is similar to polynomials. Instead of powers, consider conveniently combined roots of powers of the unknown variable  $x$ .

An integral domain  $\mathbb{K}$  consists of finite expressions of the form

$$a_0 + \sum_{i=1}^k a_i (x^{m_i})^{\frac{1}{2^{n_i}}},$$

where  $a_0, \dots, a_k \in Z, m_i, n \in Z_{>0}$ . The multiplication and addition is defined as with polynomials assuming the standard behaviour of rational powers of  $x$ . Then, the only units in  $\mathbb{K}$  are  $\pm 1$ , and all elements with  $a_0 = 0$  are reducible, but the expression  $x$ , for example, cannot be expressed as a product of irreducible elements. There are simply very few irreducible elements in  $\mathbb{K}$ .

#### 12.3.6. Euclidean division and roots of polynomials.

The fundamental tool for the discussion of divisibility, common divisors, etc. in the ring of integers  $Z$  is the procedure of division with remainder, and the Euclidean algorithm for the greatest common divisor. These procedures can be generalized.

Consider univariate polynomials  $a_k x^k + \dots + a_0$  over a general integral domain  $\mathbb{K}$ .  $a_k x^k$  is called the *leading monomial*, while  $a_k$  is the *leading coefficient*.

**Lemma** (An algorithm for division with remainder). Let  $\mathbb{K}$  be an integral domain and  $f, g \in \mathbb{K}[x]$  polynomials,  $g \neq 0$ .



**12.F.48.** For each of the following permutations, decide whether the subgroup it generates is normal in the corresponding group:

- $(1, 2, 3)$  in  $\mathbb{S}_3$ ,
- $(1, 2, 3, 4)$  in  $\mathbb{S}_4$ ,
- $(1, 2, 3)$  in  $\mathbb{A}_4$

For the last case, find the right cosets of  $\mathbb{A}_4$  by the considered subgroup. Find all  $n \geq 3$  for which the subset of all cycles of length  $n$  together with the identity is a subgroup of  $\mathbb{S}_n$ . Show that if this is so, then it is even a normal subgroup.

**Solution.**

- It is a normal subgroup of  $A_3$ .
- It is not a normal subgroup ( $(1, 2) \circ (1, 3) \circ (2, 4) \circ (1, 2) = (4, 1) \circ (2, 3)$ ).
- It is not a normal subgroup. The right cosets are

$$\begin{aligned} &\{(1, 2, 4), (2, 4, 3), (1, 3) \circ (2, 4)\}, \\ &\{(1, 4, 2), (1, 4, 3), (1, 4) \circ (2, 3)\}, \\ &\{(2, 3, 4), (1, 2) \circ (3, 4), (1, 3, 4)\}, \\ &\{\text{id}, (1, 2, 3), (1, 3, 2)\}. \end{aligned}$$

The mentioned subset is a subgroup only for  $n = 3$ . In this case, it is the alternating group  $A_3$  of all even permutations in  $\mathbb{S}_3$ , which is a normal subgroup. (For greater value of  $n$ , we can find two cycles of length  $n$  whose composition is neither a cycle of length  $n$  nor the identity.)  $\square$

**12.F.49.** Find the subgroup of  $\mathbb{S}_6$  that is generated by the permutations  $(1, 2) \circ (3, 4) \circ (5, 6)$ ,  $(1, 2, 3, 4)$ , and  $(5, 6)$ . Is this subgroup normal? If so, describe the set of (two-sided) cosets  $S_6/H$ .

**Solution.** First of all, note that all of the generating permutations lie in the subgroup  $\mathbb{S}_4 \times \mathbb{S}_2 \subseteq S_6$  (considering the natural inclusion of  $\mathbb{S}_4 \times \mathbb{S}_2$ , i. e., for  $s \in \mathbb{S}_4 \times \mathbb{S}_2$ , the restriction of  $s$  to  $\{1, 2, 3, 4\}$  is a permutation on this set and so is the restriction of  $s$  to  $\{5, 6\}$ ). This means that the group they generate is also a subgroup of  $\mathbb{S}_4 \times \mathbb{S}_2$ . Moreover, since there is  $(5, 6)$  among the generators, we can see that the subgroup is of the form  $H \times \mathbb{S}_2$ , where  $H \subseteq \mathbb{S}_4$ . Thus, it suffices to describe  $H$ . This group is generated by the elements  $(1, 2) \circ (3, 4)$  and  $(1, 2, 3, 4)$  (projection of the generators on  $\mathbb{S}_4$ ). We have

$$\begin{aligned} (1, 2, 3, 4)^2 &= (1, 3) \circ (2, 4), \\ (1, 2, 3, 4)^3 &= (4, 3, 2, 1), \\ (1, 2, 3, 4)^4 &= \text{id}, \\ [(1, 2) \circ (3, 4)]^2 &= \text{id}, \\ [(1, 2) \circ (3, 4)] \circ (1, 2, 3, 4) &= (2, 4), \\ (1, 2, 3, 4) \circ [(1, 2) \circ (3, 4)] &= (1, 3), \\ [(1, 2) \circ (3, 4)] \circ (4, 3, 2, 1) &= (1, 3), \\ (4, 3, 2, 1) \circ [(1, 2) \circ (3, 4)] &= (2, 4), \\ [(1, 2) \circ (3, 4)] \circ [(1, 3) \circ (2, 4)] &= (1, 4) \circ (2, 3), \\ [(1, 3) \circ (2, 4)] \circ [(1, 2) \circ (3, 4)] &= (1, 4) \circ (2, 3), \end{aligned}$$

Then, there exists an  $a \in \mathbb{K}$ ,  $a \neq 0$ , and polynomials  $q$  and  $r$  such that  $af = qg + r$ , where either  $r = 0$  or  $\deg r < \deg g$ . Moreover, if  $\mathbb{K}$  is a field or if the leading coefficient of  $g$  is one, and if the choice  $a = 1$  is made, then the polynomials  $q$  and  $r$  exist and they are unique.

**PROOF.** If  $\deg f < \deg g$  or  $f = 0$ , then the choice  $a = 1$ ,  $q = 0$ ,  $r = f$ , satisfies all the conditions. If  $g$  is constant, set  $a = g$ ,  $q = f$ ,  $r = 0$ . Continue by induction on the degree of  $f$ . Suppose  $\deg f \geq \deg g > 0$ , and write

$$f = a_0 + \dots + a_n x^n, \quad g = b_0 + \dots + b_m x^m.$$

Either  $b_m f - a_n x^{n-m} g = 0$  or  $\deg(b_m f - a_n x^{n-m} g) < \deg f$ . In the former case, the proof is finished. In the latter case, it follows by the induction hypothesis that there exist  $a', q', r'$  satisfying

$$a' (b_m f - a_n x^{n-m} g) = q' g + r'$$

and either  $r' = 0$  or  $\deg r' < \deg g$ . This means that

$$a' b_m f = (q' + a' a_n x^{n-m}) g + r'.$$

If  $b_m = 1$  or  $\mathbb{K}$  is a field, then the induction hypothesis can be used to choose  $a' = 1$  and then  $q', r'$  are unique. In this case,

$$b_m f = (q' + a_n x^{n-m}) g + r'.$$

If  $\mathbb{K}$  is a field, then this equation can be multiplied by  $b_m^{-1}$ .

Assume that there is another solution  $f = q_1 g + r_1$ . Then,  $0 = f - f = (q - q_1)g + (r - r_1)$  and either  $r = r_1$ , or  $\deg(r - r_1) < \deg g$ . In the former case, it follows that  $q = q_1$  as well, since there are no zero divisors in  $\mathbb{K}[x]$ . Let  $ax^s$  be the term of the highest degree in  $q - q_1 \neq 0$  (it must exist). Then, its product with the term of the highest degree in  $g$  must be zero (since the term of the highest degree is just the product of the terms of the highest degrees and there are no other terms of this degree there). However, this means that  $a = 0$ . Since  $ax^s$  is the non-zero term with the highest degree,  $q - q_1$  contains no non-zero monomials, so it equals zero. But then  $r = r_1$ .  $\square$

The procedure for the Euclidean division can be used to discuss the roots of a polynomial.

Consider a polynomial  $f \in \mathbb{K}[x]$ ,  $\deg f > 0$ , and divide it by the polynomial  $x - b$ ,  $b \in \mathbb{K}$ . Since the leading coefficient is one, the algorithm produces a unique result. It follows that there are unique polynomials  $q$  and  $r$  which satisfy  $f = q(x - b) + r$ , where  $r = 0$  or  $\deg r = 0$ , i.e.  $r \in \mathbb{K}$ . This means that the value of the polynomial  $f$  at  $b \in \mathbb{K}$  equals  $f(b) = r$ . It follows that the element  $b \in \mathbb{K}$  is a root of the polynomial  $f$  if and only if  $(x - b)|f$ . Since division by a polynomial of degree one decreases the degree of the original polynomial by at least one, the following proposition is proved:

**Corollary.** Every polynomial  $f \in \mathbb{K}[x]$  has at most  $\deg f$  roots, including multiplicities. In particular, polynomials over an infinite integral domain define the same mapping  $\mathbb{K} \rightarrow \mathbb{K}$  if and only if they are the same polynomial.

$$[(1, 2) \circ (3, 4)] \circ (4, 2) = (1, 2, 3, 4),$$

$$(1, 3) \circ (4, 2) = (1, 3) \circ (2, 4).$$

Now, we can note that the generating permutations  $(1, 2, 3, 4)$  and  $(1, 2) \circ (3, 4)$  are symmetries of a square on vertices  $1, 2, 3, 4$ . Therefore, they cannot generate more than 8-element  $D_4$ , which has already happened. This means that no more permutations can be obtained by further compositions. Thus, the subgroup  $H \subseteq S_4$  has 8 elements (which is possible by Lagrange's theorem, since 8 divides 24).

$$H = \{ \text{id}, (1, 2, 3, 4), (1, 3) \circ (2, 4), (4, 3, 2, 1), (1, 2) \circ (3, 4), (1, 3), (2, 4), (1, 4) \circ (2, 3) \}.$$

Altogether, the examined subgroup in  $S_6$  has 16 elements: for each  $h \in H$ , it contains  $(h, \text{id})$  and  $(h, (56))$ .

□

**12.F.50.** Find the subgroup in  $S_4$  that is generated by the permutations  $(1, 2) \circ (3, 4)$ ,  $(1, 2, 3)$ .

**Solution.** Since both the generating permutations are even, they can generate only even permutations. Thus, the examined group is a subgroup of the alternating group  $A_4$  of all even permutations. We have

$$[(1, 2) \circ (3, 4)]^2 = \text{id},$$

$$(1, 2, 3)^2 = (3, 2, 1),$$

$$[(1, 2) \circ (3, 4)] \circ (1, 2, 3) = (2, 4, 3),$$

$$(1, 2, 3) \circ [(1, 2) \circ (3, 4)] = (1, 3, 4),$$

$$[(1, 2) \circ (3, 4)] \circ (3, 2, 1) = (3, 1, 4),$$

$$(3, 2, 1) \circ [(1, 2) \circ (3, 4)] = (2, 3, 4),$$

and now, we already have seven elements of the examined subgroup of  $A_4$ , and since  $A_4$  has 12 elements and the order of a subgroup must divide that of the group, it is clear that the subgroup is the whole  $A_4$ . □

**12.F.51.** Find all subgroups of the group of invertible 2-by-2 matrices over  $\mathbb{Z}_2$  (with matrix multiplication). Is any of them normal?

**Solution.** In exercise 12.E.1, we built the table of the operation in this group. By Lagrange's theorem (12.4.10), the order of any subgroup must divide the order of the group, which is six. Thus, besides the trivial subgroup  $\{A\}$  and the entire group, each subgroup must have two or three elements. In a 2-element subgroup, the non-trivial element must be self-inverse, which is also sufficient for the subset to be a subgroup. We thus get the subgroups  $\{A, B\}$ ,  $\{A, C\}$ ,  $\{A, F\}$ , which are not normal, as can be easily verified. The identity is  $A$ . Since  $B, C, F$  have order 2, they cannot be in a 3-element subgroup. Thus, the only remaining possibility is  $P = \{A, D, E\}$ , which is indeed a subgroup. Moreover, checking the conjugations  $BDB = E$ ,  $CDC = E$ ,  $FD F = E$  (whence it follows that  $BEB = D$ ,  $CEC = D$ ,  $FEF = D$ ), we find out that this subgroup is normal. □

Indeed, if two polynomials over an integral domain define the same mapping  $\mathbb{K} \rightarrow \mathbb{K}$ , then their difference has any element of  $\mathbb{K}$  as a root. This means that if their difference is not the zero polynomial, then  $\mathbb{K}$  has at most as many elements as the maximum of the degrees of the polynomials in question.

**12.3.7. Multiple roots and derivatives.** In the continuous modelling, we mostly work over infinite integral domains  $\mathbb{K}$  and so we identify the algebraic expressions for the polynomials with the mappings. However the properties of differentiation on polynomials are of purely algebraic character, in general:

The differentiation of polynomials over real or complex numbers is an algebraic operation which make sense for all commutative rings  $\mathbb{K}$  and it still satisfies the Leibniz rule:



#### DERIVATIVE OF POLYNOMIALS

Consider polynomials  $f(x) = a_0 + a_1x + \dots + a_nx^n$ ,  $g(x) = b_0 + b_1x + \dots + b_mx^m$  be polynomials of orders  $n$  and  $m$  over an commutative ring  $\mathbb{K}$ . The derivative  $' : f(x) \mapsto f'(x) = a_1 + a_2x + \dots + na_nx^{n-1}$  respects the addition of polynomials and their multiplication by the elements in  $\mathbb{K}$ . Moreover it satisfies the *Leibniz rule*

$$(1) \quad (f(x)g(x))' = f'(x)g(x) + f(x)g'(x).$$

While claim on the additive structure is obvious, let us check the Leibniz rule:

$$f(x) \cdot g(x) = \sum_{k=0}^{m+n} c_k x^k, \quad c_k = \sum_{i+j=k} a_i b_j$$

and thus, expanding  $f'(x) \cdot g(x) + f(x) \cdot g'(x)$  yields exactly the expression for the derivative of the product  $\sum_{k=1}^{m+n} k c_k x^{k-1}$ .

In particular, the derivative is not a homomorphism  $\mathbb{K}[x] \rightarrow \mathbb{K}[x]$  of the ring of polynomials, in view of (1). In much more general context, the homomorphisms of the additive structure of a ring satisfying the Leibniz rule (1) are called *derivations*. For polynomial rings, we see inductively that the only derivation there is our operation  $'$ .

Differentiation can be exploited easily for discussing multiple roots of polynomials.

Consider a polynomial  $f(x) \in \mathbb{K}[x]$  over an integral domain  $\mathbb{K}$ , with root  $c \in \mathbb{K}$  of multiplicity  $k$ . Thus, in view of the division of polynomials discussed in the previous paragraph 12.3.6,

$$f(x) = (x - c)^k g(x),$$

with the unique polynomial  $g, g(c) \neq 0$ . Differentiating  $f(x)$  and applying the Leibniz rule we obtain

$$f'(x) = k(x - c)^{k-1} g(x) + (x - c)^k g'(x)$$

$$= (x - c)^{k-1} (k g(x) + (x - c) g'(x)).$$

Clearly the polynomial  $h(x) = k g(x) + (x - c) g'(x)$  does not admit  $c$  as root, i.e.  $h(c) = k g(c) \neq 0$ .

**12.F.52.** Find all subgroups of the group  $(\mathbb{Z}_{10}, +)$ .

**Solution.** The subgroups are isomorphic to  $(\mathbb{Z}_d, +)$ , Where  $d|10$ , i. e.,  $\{0\} \cong \mathbb{Z}_1$ ,  $\{0, 5\} \cong \mathbb{Z}_2$ ,  $\{0, 2, 4, 6, 8\} \cong \mathbb{Z}_5$ , and  $\mathbb{Z}_{10}$ .  $\square$

**12.F.53.** Find the orders of the elements 2, 4, 5 in  $(\mathbb{Z}_{35}^\times, \cdot)$  and in  $(\mathbb{Z}_{35}, +)$ .

**Solution.** By definition, the order of  $x$  in the group  $(\mathbb{Z}_{35}^\times, \cdot)$  is the least positive integer  $k$  such that  $x^k \equiv 1 \pmod{35}$ . By Euler's theorem, the order of  $x = 2$  and  $x = 4$  is  $k \leq \varphi(35) = 24$ . Computing the corresponding modular powers, we find out that the order of  $x = 2$  is 12. Hence it immediately follows that the order of  $x = 4$  is 6. The number  $x = 5$  does not lie in the group  $(\mathbb{Z}_{35}^\times, \cdot)$ . Specifically, we have (modulo 35):

$x$	$x^2$	$x^3$	$x^4$	$x^5$	$x^6$	$x^7$	$x^8$	$x^9$	$x^{10}$	$x^{11}$	$x^{12}$	
2	4	8	16	32	29	23	11	22	9	18	1	
4	16	29	11	9	1							

In the group  $(\mathbb{Z}_{35}, +)$ , the order of  $x$  is the least positive integer  $k$  such that  $k \cdot x \equiv 0 \pmod{35}$ . This can be calculated simply as  $k = \frac{35}{\gcd(35, x)}$ . Therefore, the order of 2 and 4 is 35, while the order of 5 is 7.  $\square$

**12.F.54.** Find all finite subgroups of the group  $(\mathbb{R}^*, \cdot)$ <sup>1</sup>

**Solution.** If a given subgroup of the group  $(\mathbb{R}^*, \cdot)$  contains an element  $a$ ,  $|a| \neq 1$ , then the elements  $a, a^2, a^3, \dots$  form an infinite geometric progression of pairwise distinct elements all of which must lie in the considered subgroup, so it is infinite. Thus, a finite subgroup may contain only the numbers 1 and  $-1$ , which means that there are two finite subgroups:  $\{1\}, \{-1, 1\}$ .  $\square$

**12.F.55.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism, and if so, find its kernel. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{Z}_4 \times \mathbb{Z}_3 \rightarrow \mathbb{Z}_{12}, \varphi([a]_4, [b]_3) = [a - b]_{12}$ ,
- ii)  $\varphi : \mathbb{Z}_4 \times \mathbb{Z}_3 \rightarrow \mathbb{Z}_{12}, \varphi([a]_4, [b]_3) = [6a + 4b]_{12}$ ,
- iii)  $\varphi : \mathbb{Z}_4 \times \mathbb{Z}_3 \rightarrow \mathbb{Z}_{12}, \varphi([a]_4, [b]_3) = [0]_{12}$ .

**Solution.**

- i) Not a mapping. For instance, if we take two representatives  $([6]_4, [1]_3) = ([2]_2, [1]_3)$  of the same element in  $\mathbb{Z}_4 \times \mathbb{Z}_3$ , then we get  $\varphi([6]_4, [1]_3) = [5]_{12}$  and  $\varphi([2]_4, [1]_3) = [1]_{12}$ , so this is not a correct definition of a mapping.
- ii) A homomorphism, neither injective, nor surjective. Its kernel  $\text{Ker}(\varphi)$  is the set  $\{([2]_4, [0]_3), ([0]_4, [0]_3)\}$ .
- iii) A homomorphism, neither injective, nor surjective. Its kernel is the entire group  $\mathbb{Z}_4 \times \mathbb{Z}_3$ .  $\square$

<sup>1</sup>The group of all invertible elements for  $\mathbb{R}$  and  $\mathbb{C}$  is denoted by  $\mathbb{R}^*$  and  $\mathbb{C}^*$ , respectively, and by  $\mathbb{Z}_n^\times$  for  $\mathbb{Z}_n$ .

On the contrary, if  $c$  is a joint root of  $f$  and its derivative, then, reading the previous computation backwards (with  $k = 1$ ), we see the  $c$  is also a root of  $g(x)$ , and thus a root with multiplicity at least two of  $f$ . Inductively, we arrive at the following very useful claim:

**Proposition.** A polynomial  $f(x)$  over an integral domain  $\mathbb{K}$  admits the root  $c \in \mathbb{K}$  of multiplicity  $k$ , if and only if  $c$  is the root of  $f'(x)$  of multiplicity  $k - 1$ .

**12.3.8. The greatest common divisor.** Consider a polynomial ring  $\mathbb{K}[x]$  over an integral domain  $\mathbb{K}$ . A polynomial  $h \in \mathbb{K}[x]$  is called the *greatest common divisor* of polynomials  $f$  and  $g \in \mathbb{K}[x]$  if and only if the following hold:

- $h|f$  and  $h|g$ ,
- for any  $k$ , if both  $k|f$  and  $k|g$ , then  $k|h$ .

As a direct corollary of the existence of an algorithm for unique division with remainder, there is the very important *Bezout's identity* (it is proved using the Euclidean division similarly as in the case of the integers in Chapter 11).

**Theorem.** Let  $\mathbb{K}$  be a field and  $f, g \in \mathbb{K}[x]$ . Then, there exists a greatest common divisor  $h$  of the polynomials  $f$  and  $g$ . The polynomial  $h$  is unique up to a multiple by a non-zero scalar. In addition, there exist polynomials  $A, B \in \mathbb{K}[x]$  such that  $h = Af + Bg$ .

**PROOF.** The polynomials  $h, A, B$  can be constructed directly using the Euclidean algorithm. Continue dividing with remainder (since  $\mathbb{K}$  is a field, there is always a unique way to do this; see the above lemma 12.3.6):

$$\begin{aligned} f &= q_1g + r_1, \\ g &= q_2r_1 + r_2, \\ r_1 &= q_3r_2 + r_3, \\ &\vdots \\ r_{p-1} &= q_{p+1}r_p + 0. \end{aligned}$$

In this procedure, the degrees of the polynomials  $r_i$  are strictly decreasing; hence the equality from the last line must occur (for some  $p$ ), and this says that  $r_p|r_{p-1}$ . It follows from the line above that  $r_p|r_{p-2}$  etc.. Continue sequentially up to the first and second lines, to obtain  $r_p|g$  and  $r_p|f$ .

If  $h|f$  and  $h|g$ , then the same equalities imply that  $h$  divides all the  $r_i$ . In particular, it divides  $r_p$ . In this way, the greatest common divisor  $h = r_p$  of the polynomials  $f$  and  $g$  is obtained.

Substitute upwards, starting with the last equation:

$$\begin{aligned} h &= r_p = r_{p-2} - q_p r_{p-1} = \\ &= r_{p-2} - q_p(r_{p-3} - q_{p-1}r_{p-2}) = \\ &= -q_p r_{p-3} + (1 + q_{p-1}q_p)r_{p-2} = \\ &= -q_p r_{p-3} + (1 + q_p q_{p-1})(r_{p-4} - q_{p-2}r_{p-3}) = \\ &\vdots \end{aligned}$$

**12.F.56.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism, and if so, find its kernel. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{Z}_4 \rightarrow \mathbb{C}^*, \varphi([a]_4) = i^a,$
- ii)  $\varphi : \mathbb{Z}_5 \rightarrow \mathbb{C}^*, \varphi([a]_5) = i^a,$
- iii)  $\varphi : \mathbb{Z}_4 \rightarrow \mathbb{C}^*, \varphi([a]_4) = (-1)^a,$
- iv)  $\varphi : \mathbb{Z} \rightarrow \mathbb{C}^*, \varphi(a) = i^a.$

**Solution.**

- i) We have  $\varphi([a]_4 + [b]_4) = i^{a+b} = i^a \cdot i^b = \varphi([a]) \cdot \varphi([b]),$  and  $\varphi([4]) = i^4 = 1,$  which means that if  $[c]_4 = [d]_4,$  i. e.,  $c = d + 4k, k \in \mathbb{Z},$  then  $\varphi([c]_4) = i^c = i^{d+4k} = i^d = \varphi([d]_4),$  so the mapping is a well-defined homomorphism. It is injective (which is equivalent to saying that  $\text{Ker}(\varphi) = \{[0]_4\}$ ), but it is clearly not surjective.
- ii) Not a mapping since we have  $[0]_5 = [5]_5$  and  $\varphi([0]_5) = i^0 = 1$  but  $\varphi([5]_5) = i^5 = i.$
- iii) Is a homomorphism, neither injective (we have  $-1 = \varphi(1) = \varphi(3) = (-1)^3 = -1$ ), nor surjective. The kernel is  $\text{Ker}(\varphi) = \{[0]_4, [2]_4\}.$
- iv) Is a homomorphism, neither injective nor surjective. The kernel is  $\text{Ker}(\varphi) = 4\mathbb{Z} = \{4k \mid k \in \mathbb{Z}\}.$

**12.F.57.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{Q}^* \rightarrow \mathbb{Q}^*, \varphi\left(\frac{p}{q}\right) = \frac{q}{p}$
- ii)  $\varphi : \mathbb{Q}^* \rightarrow \mathbb{Q}^*, \varphi\left(\frac{p}{q}\right) = \frac{p^2}{q^2}$
- iii)  $\varphi : \mathbb{Q}^* \rightarrow \mathbb{Q}^*, \varphi\left(\frac{p}{q}\right) = \frac{p^2+q^2}{pq}$

**12.F.58.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{C} \rightarrow \mathbb{R}, \varphi(a + bi) = a + b,$
- ii)  $\varphi : \mathbb{C} \rightarrow \mathbb{R}, \varphi(a + bi) = a,$
- iii)  $\varphi : \mathbb{C}^* \rightarrow \mathbb{R}^*, \varphi(a + bi) = a^2 + b^2,$
- iv)  $\varphi : \mathbb{C}^* \rightarrow \mathbb{R}^*, \varphi(c) = 2|c|,$
- v)  $\varphi : \mathbb{C}^* \rightarrow \mathbb{R}^*, \varphi(c) = |c|^3,$
- vi)  $\varphi : \mathbb{C}^* \rightarrow \mathbb{R}^*, \varphi(c) = 1/|c|.$

**12.F.59.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathcal{GL}_2(\mathbb{R}) \rightarrow \mathbb{R}^*, \varphi(A) = |A|$
- ii)  $\varphi : \mathcal{GL}_2(\mathbb{R}) \rightarrow \mathbb{R}^*, \varphi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = a^2 + b^2.$
- iii)  $\varphi : \mathcal{GL}_2(\mathbb{R}) \rightarrow \mathbb{R}^*, \varphi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = ac + bd.$

$$= Af + Bg. \quad \square$$

**12.3.9. Unique factorization.** Now follows a very useful (and elegant) statement, the proof of which is straightforward, yet it requires many technical details (and it also concerns the so called field of rational functions). It is recommended to read the following paragraph carefully. Then maybe, at the first reading, skip the technical lemmas of the proof.

**Theorem.** Let  $\mathbb{K}$  be a unique factorization domain. Then,  $\mathbb{K}[x]$  is also a unique factorization domain.



**PROOF.** The idea of the proof is very simple. Consider a polynomial  $f \in \mathbb{K}[x]$ . If  $f$  is reducible, then  $f = f_1 \cdot f_2,$  where neither of the polynomials  $f_1, f_2 \in \mathbb{K}[x]$  is a unit. Moreover, assume for a while that if  $f$  is divisible by an irreducible polynomial  $h,$  then so is  $f_1$  or  $f_2.$

If this is always the case, this procedure can be applied step by step to reach a unique factorization. If  $f_1$  is further reducible, then  $f_1 = g_1 \cdot g_2,$  where  $g_1, g_2$  are not units, and either both the polynomials  $g_1$  and  $g_2$  have degree less than that of  $f,$  or the number of irreducible factors in the leading terms of  $g_1$  and  $g_2$  decreases (for instance, over the integers  $\mathbb{Z}, 2x^2 + 2x + 2 = 2(x^2 + x + 1)$ ). After a finite number of steps, a factorization  $f = f_1 \dots f_r$  is obtained, where the polynomials  $f_1, \dots, f_r$  are irreducible.

It follows from the additional assumption that every irreducible polynomial  $h$  which divides  $f$  also divides one of  $f_1, \dots, f_r.$  Therefore, for every other factorization  $f = f'_1 f'_2 \dots f'_s,$  each factor  $f_i$  divides one of  $f'_j,$  and in this case,  $f'_j = e f_i$  for an appropriate unit  $e.$  Cancel such pairs step by step, to conclude that  $r = s$  and that the individual factors differ only by a unit multiple.  $\square$

To conclude, we have to deal with our assumption in the first paragraph of our above proof. This will require some technical preparation and we shall come back to it soon.

The direct consequence of the latter theorem for the multivariate polynomials can be formulated (due to their inductive definition):

**Corollary.** Let  $\mathbb{K}$  be a unique factorization domain. Then,  $\mathbb{K}[x_1, \dots, x_r]$  is also a unique factorization domain.

Every polynomial over a unique factorization domain can be factored in a similar way to the case of polynomials with real or complex coefficients.

In particular, this holds for polynomials over every field of scalars.

**12.3.10. Fields of fractions.** When dealing with integer calculations, it is often more advantageous to work with rational numbers and verify only at the end of the procedure that the result is an integer. This method is useful in the case of polynomials, too.





**12.F.60.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{Z}_3 \rightarrow \mathbb{A}_4, \varphi([a]_3) = (1, 2, 4) \circ (1, 3, 2)^a \circ (1, 4, 2)$
- ii)  $\varphi : \mathbb{Z}_3 \rightarrow \mathbb{A}_4, \varphi([a]_3) = (1, 2) \circ (1, 3, 2)^a$

**12.F.61.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{Z} \rightarrow \mathbb{Z}, \varphi(a) = 2a$
- ii)  $\varphi : \mathbb{Z} \rightarrow \mathbb{Z}, \varphi(a) = a + 1$
- iii)  $\varphi : \mathbb{Z} \rightarrow \mathbb{Z}, \varphi(a) = 3|a|$
- iv)  $\varphi : \mathbb{Z} \rightarrow \mathbb{Z}, \varphi(a) = 1$

**12.F.62.** For each of the following formulas, decide whether it correctly defines a mapping  $\varphi$ . If so, decide whether it is a homomorphism. Moreover, decide whether it is surjective and injective:

- i)  $\varphi : \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Q}^*, \varphi((a, b, c)) = 2^a 3^b 12^c$
- ii)  $\varphi : \mathbb{Z}_3^* \times \mathbb{Z}_5 \rightarrow \mathbb{Z}_5, \varphi((a, b)) = b^a$
- iii)  $\varphi : \mathbb{Z}_2 \times \mathbb{Z} \rightarrow \mathbb{Z}, \varphi(([a]_2, b)) = b$

**12.F.63.** Prove that there exists no isomorphism of the multiplicative group of non-zero complex numbers onto the multiplicative group of non-zero real numbers.

**Solution.** Every homomorphism must map the identity of the domain to the identity of the codomain (see 12.4.5). Thus, 1 must be mapped to itself. And what about  $-1$ ? We know that  $f(-1)^2 = f((-1)^2) = f(1) = 1$ . Therefore, the image of  $-1$  is a square root of 1. Since we are interested in bijective homomorphisms only, we must have  $f(-1) = -1$ . However, then  $f(i)^2 = f(i^2) = f(-1) = -1$ , so that  $f(i)$  is a square root of  $-1$  in  $\mathbb{R}$ ; however, no such real number exists. Therefore, no bijective homomorphism may exist.  $\square$

**Remark.** The mapping which assigns the absolute value to each non-zero complex number is a surjective homomorphism of  $\mathbb{C}^*$  onto  $\mathbb{R}^+$ .

**G. Burnside's lemma**

**12.G.1.** How many necklaces can be created from 3 black and 6 white beads? Beads of one color are indistinguishable, and two necklaces are considered the same if they can be transformed to each other by rotation and/or reflection.

Let  $\mathbb{K}$  be an integral domain. Its *field of fractions* is defined as the set of equivalence classes of the pairs  $(a, b) \in \mathbb{K} \times \mathbb{K}, b \neq 0$ . These classes are written  $\frac{a}{b}$ , and the equivalence is defined by

$$\frac{a}{b} = \frac{a'}{b'} \iff ab' = a'b.$$

Addition and multiplication is defined in terms of representatives:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd},$$

$$\frac{a}{b} \frac{c}{d} = \frac{ac}{bd}.$$

It is easily verified that this definition is correct and that the resulting structure satisfies all field axioms. In particular,  $\frac{0}{1}$  is the additive identity, and  $\frac{1}{1}$  is the multiplicative identity. If  $a \neq 0, b \neq 0$ , then  $\frac{a}{b} \frac{b}{a} = \frac{1}{1}$ . All the details of the arguments are in fact identical with the discussion of rational numbers in 1.6.6.

The field of fractions of a ring  $\mathbb{K}[x_1, \dots, x_r]$  is called the *field of rational functions* (of  $r$  variables) and denoted  $\mathbb{K}(x_1, \dots, x_r)$ . In software systems like Sage, Maple or Mathematica, all algebraic operations with polynomials are performed in the corresponding field of fractions, i.e. in the field of rational functions, usually using  $\mathbb{K} = \mathbb{Q}$ .

**12.3.11. Completion of the proof.** It remains to prove that if a polynomial  $f = f_1 f_2$  is divisible by an irreducible polynomial  $h$ , then  $h$  divides either  $f_1$  or  $f_2$  or both.



This statement is proved in the following three lemmas.

**Lemma.** Let  $\mathbb{K}$  be a unique factorization domain. Then:

- (1) If  $a, b, c \in \mathbb{K}$ ,  $a$  is irreducible and  $a|bc$ , then either  $a|b$  or  $a|c$ .
- (2) If a constant polynomial  $a \in \mathbb{K}[x]$  divides  $f \in \mathbb{K}[x]$ , then  $a$  divides all coefficients of  $f$ .
- (3) If  $a$  is an irreducible constant polynomial in  $\mathbb{K}[x]$  and  $a|fg, f, g \in \mathbb{K}[x]$ , then  $a|f$  or  $a|g$ .

**PROOF.** (1) By the assumption,  $bc = ad$  for a suitable  $d \in \mathbb{K}$ . Let  $d = d_1 \dots d_r, b = b_1 \dots b_s, c = c_1 \dots c_q$  be the factorizations to irreducible factors. This means that

$$ad_1 \dots d_r = b_1 \dots b_s c_1 \dots c_q.$$

Since  $ad$  factors in a unique way, it follows that  $a = eb_j$  or  $a = ec_i$  for a suitable unit  $e$ .

(2) Let  $f = b_0 + b_1 x + \dots + b_n x^n$ . Since  $a|f$ , there must exist a polynomial  $g = c_0 + c_1 x + \dots + c_k x^k$  such that  $f = ag$ . Hence it immediately follows that  $k = n, ac_0 = b_0, \dots, ac_n = b_n$ .

(3) Consider  $f, g \in \mathbb{K}[x]$  as above and suppose that  $a$  divides neither  $f$  nor  $g$ . By the previous claim, there exists an  $i$  such that  $a$  does not divide  $b_i$ , and there exists a  $j$  such that  $a$  does not divide  $c_j$ . Choose the least such  $i$  and  $j$ . The coefficient at  $x^{i+j}$  of the polynomial  $fg$  is  $b_0 c_{i+j} + b_1 c_{i+j-1} + \dots + b_{i+j} c_0$ . By choice,  $a$  divides all of  $b_0 c_{i+j}, \dots, b_{i-1} c_{j+1}, b_{i+1} c_{j-1}, \dots, b_{i+j} c_0$ . At the same

**Solution.** Let us assume the necklace as coloring of the vertices of a regular 9-gon. Let  $S$  denote the set of all such colorings. Since each coloring is determined by the positions of the 3 black beads, we get that  $S$  has  $\binom{9}{3} = 84$  elements.

We know that the group of symmetries is  $D_9$ , which contains 9 rotations (including the identity) and 9 reflections. Two colorings are the same if they lie in the same orbit of the action of  $D_9$  on the set  $S$ . Thus, we are interested in the number of orbits (let us denote it  $N$ ). In order to find  $N$ , it suffices to compute the sizes of  $S_g$  for all elements  $g$  of  $D_9$ :

The identity is the only element of order 1, we have  $|S_{\text{id}}| = 84$ , so the contribution to the sum is 84.

There are 9 reflections  $g$ , each of order 2. Clearly, we have  $|S_g| = 4$ , so the total contribution is  $4 \cdot 9 = 36$ .

There are 2 rotations  $g$  by  $2\pi/3$  or  $4\pi/3$ , both of order 3, and  $|S_g| = 3$ . Their contribution is 6.

Finally, there are 6 rotations of order 9, and no coloring is kept unchanged by them, so they do not contribute to the sum.

Altogether, we get by the formula of Burnside's lemma that

$$N = \frac{1}{|D_9|} \sum_{g \in D_9} |S_g| = \frac{126}{18} = 7.$$

Draw the seven necklaces! □

**12.G.2.** Find the number of colorings of a 3-by-3 grid with three colors. Two colorings are considered the same if they can be transformed to each other by rotation and/or reflection.

**Solution.** The group of symmetries of the table is the same as for a square, i. e., it is the dihedral group  $D_4$ . Without any identification, there are clearly  $3^9$  colorings of the table. Now, the group  $G = D_4$  acts on these colorings. For each symmetry  $g \in G$ , we find the number of colorings that  $g$  keeps unchanged:

- $g = \text{Id}$ :  $|S_g| = 3^9$ .
- $g$  is a rotation by  $90^\circ$  or  $270^\circ (= -90^\circ)$ : In this rotation, every corner tile is sent to an adjacent corner tile. This means that if the coloring is to be unchanged, all the corner tiles must be of the same color. Similarly, all the edge tiles must be of the same color. Then, the center tile may be of any color. Altogether, we get that there are  $3^3$  which are not changed by the considered rotations.
- $g$  is a rotation by  $180^\circ$ : There are four pairs of tiles that are sent to each other by this symmetry, which means that the two tiles of each pair must be of the same color. Then, the center tile may be of any color. Altogether, we have  $|S_g| = 3^5$ .
- $g$  is one of the four reflections: There are three pairs of tiles that are sent to each other by the reflection, so again the tiles within each pair must be of one color. The three tiles that are fixed by the reflection may be each of an arbitrary color. Altogether, we get  $|S_g| = 3^6$ .

time, it does not divide  $b_i c_j$ . Therefore, it cannot divide the coefficient. □

**12.3.12. Lemma.** Consider the field of fractions  $\mathbb{L}$  of a unique factorization domain  $\mathbb{K}$ . If a polynomial  $f$  is irreducible in  $\mathbb{K}[x]$ , then it is irreducible in  $\mathbb{L}[x]$ , too.

**PROOF.** Each coefficient  $a \in \mathbb{K}$  can be considered as an element  $\frac{a}{1} \in \mathbb{L}$ . Therefore, every non-zero polynomial  $f \in \mathbb{K}[x]$  can be considered a polynomial in  $\mathbb{L}[x]$ .

Suppose that  $f = g'h'$  for some  $g', h' \in \mathbb{L}[x]$ , where the polynomials  $g', h'$  are not units in  $\mathbb{L}[x]$  (i.e. they are not constant polynomials, since  $\mathbb{L}$  is a field). Let  $a$  be a common multiple of the denominators of the coefficients in  $g'$  and  $b$  be a common multiple of the denominators of coefficients in  $h'$ . Then,  $bh', ag' \in \mathbb{K}[x]$ , and so  $abf = (bh')(ag')$ . Let  $c$  be an irreducible factor in the factorization of  $ab$ . Then,  $c$  divides  $(bh')(ag')$ , and hence  $c$  divides  $bh'$  or  $ag'$  (by the previous lemma). This means that  $c$  can be canceled out. After a finite number of such cancellations, the conclusion is that  $f = gh$  for polynomials  $g, h \in \mathbb{K}[x]$ . Since the degrees of the polynomials are not changed, neither  $g$  nor  $h$  is constant.

Thus if  $f$  is reducible in  $\mathbb{L}[x]$ , then it is also reducible in  $\mathbb{K}[x]$ , contradicting the implication to be proved. □

**12.3.13. Lemma.** Let  $\mathbb{K}$  be a unique factorization domain and  $f, g, h \in \mathbb{K}[x]$ . Suppose that  $f$  is irreducible and  $f|gh$ . Then, either  $f|g$  or  $f|h$ .

**PROOF.** This statement is already proved in one of the previous lemmas for the case that  $f$  is a constant polynomial (i.e. an element of  $\mathbb{K}$ ).

Suppose that  $\deg f > 0$ . Then  $f$  is irreducible in  $\mathbb{L}[x]$  as well, where  $\mathbb{L}$  is the field of fractions of the ring  $\mathbb{K}$ .

Suppose that  $\mathbb{K}$  itself is a field (and as such equals its field of fractions). Moreover, suppose that  $f|gh$  and  $f$  does not divide  $g$ . The greatest common divisor of the polynomials  $g$  and  $f$  must be a constant polynomial in  $\mathbb{K}$ . Therefore, there are  $A, B \in \mathbb{K}[x]$  such that  $1 = Af + Bg$ . Hence,  $h = Afh + Bgh$ . Since  $f|gh$ , it follows that  $f|h$  as well.

Return to the general case. It follows from the assumptions that  $f|g$  or  $f|h$  in the polynomial ring  $\mathbb{L}[x]$  over the field of fractions  $\mathbb{L}$  of the ring  $\mathbb{K}$ . For instance, let  $h = kf$  in  $\mathbb{L}[x]$ , and choose an  $a \in \mathbb{K}$  so that  $ak \in \mathbb{K}[x]$ . Then,  $ah = akf$  and it must hold for every irreducible factor  $c$  of  $a$  that  $c|ak$ , because  $f$  is irreducible and not constant.

It follows that  $c$  can be canceled. After a finite number of such cancellations,  $a$  becomes a unit, i.e.  $h = k'f$  for an appropriate  $k' \in \mathbb{K}[x]$ . □

The proof of this lemma completes the proof of theorem 12.3.9.

By Burnside's lemma, the wanted number of colorings is equal to

$$\frac{1}{8} (3^9 + 2 \cdot 3^3 + 3^5 + 4 \cdot 3^6) = 2862. \quad \square$$

**12.G.3.**

- a) Find all rotational symmetries of a regular octahedron.
- b) Find the number of colorings of its sides. Two colorings are considered the same they can be transformed to each other by rotation.

**Solution.**

- a) Placing the octahedron into the Cartesian coordinate system so that pairs of adjacent vertices are on the axes and the center of the octahedron lies in the origin, then every rotational symmetry is given by which of the six vertices is on the positive  $x$ -semi-axis and which of the four adjacent vertices is on the positive  $y$ -semi-axis. Thus, the group has 24 elements. These are (besides the identity) rotations by  $\pm 90^\circ$  and  $180^\circ$  around axes going through opposite vertices, rotations by  $180^\circ$  around axes going through the centers of opposite edges, and finally rotations around  $\pm 120^\circ$  around axes going through the centers of opposite sides.
- b) Without any identifications, there are  $3^8$  colorings. For each rotational symmetry  $g$ , we compute the number of colorings that are kept unchanged by it:
  - $g$  is a rotation by  $\pm 90^\circ$  around an axis going through opposite vertices. Then,  $g$  fixes  $3^2$  colorings, and there are 6 such rotations.
  - $g$  is a rotation by  $180^\circ$  around an axis going through opposite vertices or the centers of opposite edges. Then,  $g$  fixes  $3^4$  colorings. There are  $3 + 6 = 9$  of these.
  - $g$  is a rotation by  $\pm 120^\circ$ . Then,  $g$  also fixes  $3^4$  colorings, and there are 8 such rotations.

Together with  $3^8$  for the identity, we get that the number of colorings is

$$\frac{1}{24} (3^8 + 6 \cdot 3^2 + 17 \cdot 3^4) = 333. \quad \square$$

**12.G.4.** How many necklaces can be created from 9 white, 6 red, and 3 black beads? Beads of one color are indistinguishable, and two necklaces are considered the same if they can be transformed to each other by rotation and/or reflection.

**4. Groups, rings, and fields**

As an illustration of the most abstract approach to an algebraic theory, concepts enjoying just one operation only are considered. The focus is on objects and situations where equations of the form  $a \cdot x = b$  always have a unique solution (as usual with linear equations, the objects  $a$  and  $b$  are given, while  $x$  is what is sought for). This is group theory. Note that nothing is known about the "nature" of the objects, or even what the dot stands for. The only assumption is that two objects  $a$  and  $x$  are assigned an object  $a \cdot x$ .

In a previous part of this chapter, such operations are known as addition or multiplication in rings. The concepts and vocabulary concerning such operations are now extended. Among them, numbers and transformations of the plane and space, where such "group" objects are met. Then follows the foundations of a general theory.

**12.4.1. Examples and concepts.** Let  $A$  be a set. A *binary operation* on  $A$  is defined to be any mapping  $A \times A \rightarrow A$ . The result of such an operation is often denoted

$$(a, b) \mapsto a \cdot b$$

and called the product of  $a$  and  $b$ . A set together with a binary operation is called a *groupoid* or a *magma*.

Further assumed properties of the operations are needed in order to be able to say something interesting,

**BINARY OPERATIONS AND SEMIGROUPS**

A binary operation is said to be *associative*, if and only if  $a \cdot (b \cdot c) = (a \cdot b) \cdot c$  for all  $a, b, c \in A$ .

A groupoid where the operation is associative is called a *semigroup*.

A binary operation is said to be *commutative* if and only if  $a \cdot b = b \cdot a$  for all  $a, b \in A$ .

The natural numbers  $\mathbb{N} = \{0, 1, 2, \dots\}$  together with either addition or multiplication form a groupoid. These operations are both commutative and associative. The integers  $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$  form a groupoid with any of addition, subtraction, and multiplication. Subtraction is neither associative, for example

$$(5 - 3) - 2 = 0 \neq 5 - (3 - 2) = 4,$$

nor commutative, since  $a - b = -(b - a)$ , which is in general different from  $b - a$ .



**Solution.** The group of symmetries of the necklace is the dihedral group  $D_{18}$ , which has 36 elements. It acts on the set of necklaces, where we can number each place (1 through 18), resulting in  $18!/(9!6!3!) = 4084080$  necklaces (without any identification). The only symmetries that fix a non-zero number of necklaces are rotations by  $120^\circ$  and  $240^\circ$ , reflections, and of course the identity. By Burnside's lemma, the wanted number of necklaces is equal to

$$\frac{1}{36} \left( 4084080 + 2 \cdot \binom{6}{3} \binom{3}{3} + 9 \cdot \binom{8}{4} \binom{4}{3} \right) = 113590.$$

□

**12.G.5.** How many necklaces can be created from 6 white, 6 red, and 6 black beads? Beads of one color are indistinguishable, and two necklaces are considered the same if they can be transformed to each other by rotation and/or reflection.

○

**12.G.6.** How many necklaces can be created from 8 white, 8 red, and 8 black beads? Beads of one color are indistinguishable, and two necklaces are considered the same if they can be transformed to each other by rotation and/or reflection.

○

**12.G.7.** How many necklaces can be created from 3 white and 6 black beads? Beads of one color are indistinguishable, and two necklaces are considered the same if they can be transformed to each other by rotation and/or reflection.

○

### H. Codes

**12.H.1.** Consider the  $(5, 3)$ -code over  $\mathbb{Z}_2$  generated by the polynomial  $x^2 + x + 1$ . Find all codewords as well as the generating matrix and the check matrix.

**Solution.**  $p(x) = x^2 + x + 1$ . The code words are precisely the multiples of the generating polynomial:

$$0 \cdot p, 1 \cdot p, x \cdot p, (x+1) \cdot p, x^2 \cdot p, (x^2+1) \cdot p, (x^2+x) \cdot p, (x^2+x+1) \cdot p$$

or

$$0, x^2 + x + 1, x^3 + x^2 + x, x^3 + 1, x^4 + x^3 + x^2, x^4 + x^3 + x + 1, x^4 + x, x^4 + x^2 + 1$$

### NEUTRAL ELEMENTS, INVERSES, AND GROUPS<sup>7</sup>

A *left identity* (or *left neutral element*) in a groupoid  $(A, \cdot)$  is an element  $e \in A$  such that  $e \cdot a = a$  for all  $a \in A$ . Similarly,  $e \in A$  is a *right identity* (*right neutral element*) iff for all  $a \in A$ ,  $a \cdot e = a$ . If  $e$  satisfies both these properties, it is called an *identity* (or a *neutral element*).

In a groupoid  $(A, \cdot)$  with identity  $e$ , an element  $b$  is a *left inverse* of an element  $a$  if and only if  $b \cdot a = e$ ; it is a *right inverse* of  $a$  if and only if  $a \cdot b = e$ . If  $b$  satisfies both these properties, it is called an *inverse* of  $a$ .

A *monoid*  $(M, \cdot)$  is a semigroup which has a neutral element. A *group*  $(G, \cdot)$  is a monoid where each element has an inverse.

A *commutative semigroup* is a semigroup where the operation is commutative, similarly for a *commutative monoid* or a *commutative group*. A commutative group is also often called an *Abelian group*.

Consider direct consequences of the definitions. A groupoid cannot have both a left identity and a different right identity (if it had, what would be their product equal to?). Thus, if a groupoid has a (two-sided) identity, then it is the only identity element, called *the* identity.

Similarly, in a monoid, an element  $x$  cannot have both a left inverse  $a$  and a different right inverse  $b$ , since if  $a \cdot x = x \cdot b = e$ , then also

$$a = a \cdot (x \cdot b) = (a \cdot x) \cdot b = b.$$

Note that associativity of the operation is needed here. It follows that if  $x$  has an inverse, then it is unique. It is usually denoted by  $x^{-1}$ .

As an example, consider again the subtraction on integers. This operation is not associative. There is a right identity (zero), i.e.  $a - 0 = a$  for any integer  $a$ , but it is not a left identity. There is no left identity for subtraction.

The integers are a semigroup with respect to either addition or multiplication. They form a group only with addition, since with respect to multiplication, only the integers  $\pm 1$  have an inverse.

If  $(A, \cdot)$  is a group, then any subset  $B \subseteq A$  which is closed with respect to the restriction of  $\cdot$  (i.e.  $a \cdot b \in B$  for any  $a, b \in B$ ) and forms a group with this operation is called a *subgroup*. Both conditions are essential. For instance, consider the integers as a subset of the rational numbers and the multiplication there.

Let  $G$  be a group and  $M \subset G$ . The *subgroup generated by  $M$*  is the smallest (with respect to set inclusion) subgroup of  $G$  which contains all the elements of  $M$ . Clearly, this is the intersection of all subgroups containing  $M$ .

<sup>7</sup>The name "Abelian" is in honour of a young mathematician Niels Henrik Abel. The adjective is so widely used that it is common to write it with a lower-case 'a', abelian, although it is derived from a surname.

or

00000, 11100, 01110, 10010, 00111, 11011, 01001, 10101.

The basis vectors multiplied by  $x^{5-3} = x^2$  yield mod ( $p$ ):

$$\begin{aligned} x^2 &\equiv x + 1, \\ x^3 &= x \cdot x^2 \equiv x(x + 1) = x^2 + x \equiv 1, \\ x^4 &\equiv x. \end{aligned}$$

This means that the basis vectors are encoded as follows:

$$\begin{array}{ll} 1 \mapsto x^2 + x + 1, & 100 \mapsto 11100, \\ x \mapsto x^3 + 1, & \text{i. e. } 010 \mapsto 10010, \\ x^2 \mapsto x^4 + x, & 001 \mapsto 01001. \end{array}$$

Thus, the generating matrix is

$$G = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and the check matrix is

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{pmatrix}. \quad \square$$

**12.H.2.** Consider the (5, 3)-code over  $\mathbb{Z}_2$  generated by the polynomial  $x^2 + x + 1$ . Find the generating matrix and the check matrix of the (7, 4)-code over  $\mathbb{Z}_2$  generated by the polynomial  $x^3 + x + 1$ . ○

**12.H.3.** A 7-bit message  $a_0a_1 \dots a_6$ , considered as the polynomial  $a_0 + a_1x + \dots + a_6x^6$ , is encoded using the polynomial code generated by  $x^4 + x + 1$ .

- i) Encode the message 1100011.
- ii) You have received the word 10111010001. What was the message if you assume that at most one bit was flipped?
- iii) What was the message in ii) if you assume that exactly two bits were flipped?

**Solution.** i)

$$\begin{aligned} x^4 &\equiv x + 1, \\ x^5 &\equiv x^2 + x, \\ x^9 &\equiv x^3 + x, \\ x^{10} &\equiv x^2 + x + 1, \end{aligned}$$

whence

$$\begin{aligned} 1 + x + x^5 + x^6 &\mapsto x^4 + x^5 + x^9 + x^{10} + x + 1 + x^2 + x + \\ &= x^3 + x + x^2 + x + 1 = \\ &= x^3 + x^4 + x^5 + x^9 + x^{10}. \end{aligned}$$

Thus, the code is 00011100011.



Here are a few very well known examples of groups. The rational numbers  $\mathbb{Q}$  are a commutative group with respect to addition. The integers are one of their subgroups. The non-zero rational numbers are a commutative group.

For every positive integer  $k$ , the set of all  $k$ -th roots of unity, i.e. the set  $\{z \in \mathbb{C}; z^k = 1\}$  is a finite commutative group with respect to multiplication of complex numbers. For  $k = 2$ , this is the two-element group  $\{-1, 1\}$ , both of whose elements are self-inverse. For  $k = 4$ , this is the group  $G = \{1, i, -1, -i\}$ .

The set  $\text{Mat}_n$ ,  $n > 1$ , of all square matrices of order  $n$  is a (non-commutative) monoid with respect to multiplication and a commutative group with respect to addition (see subsections 2.1.2–2.1.5).

The set of all linear mappings  $\text{Hom}(V, V)$  on a vector space is a monoid with respect to mapping composition and a commutative group with respect to addition (see subsection 2.3.12).

In every monoid, the subset of all invertible elements forms a group. In the former of the above examples, it was the group of invertible matrices. In the latter case, it was the group of linear isomorphisms of the corresponding vector space.

In previous chapters, there are several (semi)group structures, sometimes met quite unexpectedly. For example, recall various subgroups of the group of matrices or the group structure on elliptic curves.

**12.4.2. Permutation groups.** Groups and semigroups often arise as sets of mappings on a fixed set  $M$ , which are closed with respect to mapping composition.

This is easily seen on finite non-empty sets  $M$ , where every subset of invertible mappings generates a group with respect to composition.

Such a set  $M$  consisting of  $m = |M| \in \mathbb{N}$  elements allows for  $m^m$  possible mappings (each of  $m$  elements can be sent to arbitrary element of  $M$ ), and all of these mappings can be composed. Since mapping composition is associative, it is a semigroup.

If a mapping  $\alpha : M \rightarrow M$  is required to have an inverse  $\alpha^{-1}$ , then  $\alpha$  must be a bijection. Composition of two bijections yields again a bijection; hence the set  $\Sigma_m$  of all bijections on an  $m$ -element set  $M$  is a group. This is called the *symmetric group* (on  $m$  elements). It is an example of a finite group.<sup>8</sup>

The name of the group  $\Sigma_m$  brings another connection: Instead of bijections on a finite set, permutations can be viewed as the rearranging of distinguished objects. Permutations are encountered in this sense when studying determinants, for example, see subsection 2.2.1 on page 96 for a few

<sup>8</sup>It can be proved that every finite group is a subgroup of an appropriate finite symmetric group. This can be interpreted so that the groups  $\Sigma_m$  are as non-commutative and complex as possible.

ii)  $1 + x^2 + x^3 + x^4 + x^6 + x^{10}$  divide by  $x^4 + x + 1$  gives remainder  $x^2 + 1 \equiv x^8$ . Thus, the ninth bit was flipped and the original message was 1010101.

iii) Either the first and third bits were flipped ( $x^2 + 1$ ), or the fifth and sixth were ( $x^4 + x^5 \equiv x^2 + 1$ ). In the first case, the message was 1010001, while in the second case, it was 0110001.  $\square$

**12.H.4.** A 7-bit message  $a_0a_1 \dots a_6$ , considered as the polynomial  $a_0 + a_1x + \dots + a_6x^6$ , is encoded using the polynomial code generated by  $x^4 + x^3 + 1$ .

- i) Encode the message 1101011.
- ii) You have received the word 01001011101. What was the message if you assume that at most one bit was flipped?
- iii) What was the message in ii) if you assume that exactly two bits were flipped?

**Solution.** i)

$$\begin{aligned} x^4 &\equiv x^3 + 1, \\ x^5 &\equiv x^3 + x + 1, \\ x^7 &\equiv x^2 + x + 1, \\ x^9 &\equiv x^2 + 1, \\ x^{10} &\equiv x^3 + x, \end{aligned}$$

thus we get

$$\begin{aligned} 1 + x + x^3 + x^5 + x^6 &\mapsto x^4 + x^5 + x^7 + x^9 + x^{10} + x^3 + \\ &+ 1 + x^3 + x + 1 + x^2 + x + 1 + x^2 + 1 + x^3 + x = \\ &= x^3 + x^4 + x^5 + x^7 + x^9 + x^{10} + x^3 + x. \end{aligned}$$

Therefore, the code is  $\underbrace{0101}_{\text{redundancy}} \underbrace{1101011}_{\text{message}}$ .

ii)  $x + x^4 + x^6 + x^7 + x^8 + x^{10}$  divided by  $x^4 + x^3 + 1$  gives remainder  $x^2 + x + 1 \equiv x^7$ . Thus, the eighth bit was flipped, and the original message was 1010101.

iii) Either the second and tenth bit were flipped ( $x + x^9 \equiv x^2 + x + 1$ ), or the fourth and seventh ( $x^3 + x^6 \equiv x^2 + x + 1$ ), or the fifth and ninth ( $x^4 + x^8 \equiv x^2 + x + 1$ ). The respective messages are 00001011111, 01011010101, and 01000011001.  $\square$

**12.H.5.** Consider the (15, 11)-code generated by the polynomial  $1 + x^3 + x^4$ . We have received the word 011101110111001. Find the original 11-bit message, provided exactly one bit was flipped.

**Solution.** The word is a codeword if and only if it is divisible by the generating polynomial  $1 + x^3 + x^4$ . The received word corresponds to the polynomial  $x + x^2 + x^3 + x^5 + x^6 + x^7 + x^9 + x^{10} + x^{11} + x^{14}$ . When divided by  $1 + x^3 + x^4$ , it leaves remainder  $x + 1$ . This means that an error has occurred. If we assume that only one bit was flipped, then there must be a power of  $x$  which is equal to this remainder modulo  $1 + x^3 + x^4$ . Thus, we compute  $x^4 \equiv x^3 + 1$ ,  $x^5 \equiv x^3 + x + 1$ ,  $\dots$ ,  $x^{12} \equiv x + 1$  and find out that the thirteenth bit was flipped, and the original message was 01110111101.

elementary results. Let us briefly recollect them now in view of the general concepts of groups and their homomorphisms.

What the operation in this group looks like needs more thought. In the case of a (small) finite group, build a complete table of the operation results for all pairs of operands. Considering the group  $\Sigma_3$  on numbers  $\{1, 2, 3\}$  and denoting the particular permutations by the ordering of the numbers (not to be confused with the notation for cycles!)

$$\begin{aligned} a &= (1, 2, 3), \quad b = (2, 3, 1), \quad c = (3, 1, 2), \\ d &= (1, 3, 2), \quad e = (3, 2, 1), \quad f = (2, 1, 3), \end{aligned}$$

then the composition is given by the following table:

$\cdot$	$a$	$b$	$c$	$d$	$e$	$f$
$a$	$a$	$b$	$c$	$d$	$e$	$f$
$b$	$b$	$c$	$a$	$f$	$d$	$e$
$c$	$c$	$a$	$b$	$e$	$f$	$d$
$d$	$d$	$e$	$f$	$a$	$b$	$c$
$e$	$e$	$f$	$d$	$c$	$a$	$b$
$f$	$f$	$d$	$e$	$b$	$c$	$a$

Note that there is a fundamental difference between the permutations  $a, b, c$  and the other three. The former three form a *cycle*, generated by either  $b$  or  $c$ :

$$b^2 = c, \quad b^3 = a, \quad c^2 = b, \quad c^3 = a.$$

It follows that these three permutations form a commutative subgroup. Here (as well as in the whole group),  $a$  is the neutral element and  $b$  and  $c$  are inverses of each other. Therefore, this subgroup is the same as the group  $\mathbb{Z}_3$  of residue classes modulo 3, or as the group of third roots of unity.

The other three permutations are self-inverse, which means that any one of them together with the identity  $a$  create a subgroup, the same one as  $\mathbb{Z}_2$ . Further,  $b$  and  $c$  are *elements of order 3*, i.e. the third power is the first one equal to the identity  $a$ , while  $d, e$ , and  $f$  are of order 2.

Since the table is not symmetric with respect to the main diagonal, the composition  $\cdot$  is not commutative.

Other permutation groups  $\Sigma_m$  of finite  $m$ -element sets behave similarly. Each permutation  $\sigma$  partitions the set  $M$  into a disjoint union of maximal invariant subsets, which are obtained by taking unprocessed elements  $x \in M$  step by step and putting all iteration results  $\sigma^k(x)$ ,  $k = 1, 2, \dots$ , into the class  $M_x$  until  $\sigma^k(x) = x$ .

Each permutation is obtained as a composition of the cycles, which behave as the identity outside  $M_x$  and as  $\sigma$  on  $M_x$ . If the elements of  $M_x$  are numbered as  $(1, 2, \dots, |M_x|)$  so that  $i$  corresponds to  $\sigma^i(x)$ , then the permutation is simply a one-place shift in the cycle (i.e. the last element is mapped back to the first one). Hence the name *cycle*. These cycles commute, so it does not matter in which order the permutation  $\sigma$  is composed from them. (Of course, if we pick arbitrary two cycles on  $M$ , they do not have to commute.)

The simplest cycles are one-element fixed points of  $\sigma$  and two-element subsets  $(x, \sigma(x))$ , where  $\sigma(\sigma(x)) = x$ . The latter are called *transpositions*. Since every cycle can be composed from transpositions of adjacent elements (just let the

Let us look at the exercise more thoroughly. Computing all powers of  $x$ , we obtain

$$\begin{aligned} x^4 &\equiv x^3 + 1, \\ x^5 &\equiv x^3 + x + 1, \\ x^6 &\equiv x^3 + x^2 + x + 1, \\ x^7 &\equiv x^2 + x + 1, \\ x^8 &\equiv x^3 + x^2 + x, \\ x^9 &\equiv x^2 + 1, \\ x^{10} &\equiv x^3 + x, \\ x^{11} &\equiv x^3 + x^2 + 1, \\ x^{12} &\equiv x + 1, \\ x^{13} &\equiv x^2 + x, \\ x^{14} &\equiv x^3 + x^2, \end{aligned}$$

so the generating matrix is

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We can verify that multiplication by 01110111101 yields the codeword 011101110111101, which differs from the received word 011101110111001 exactly in the thirteenth bit  $\square$

Now, we begin to efficiently use the check matrix.

**12.H.6.** Find the generating matrix and the check matrix of the  $(7, 2)$ -code (i. e., there are 2 bits of the message and 5 redundant bits) generated by the polynomial  $x^5 + x^4 + x^2 + 1$ . Decode the received word 0010111 (i. e., find the message that was sent) assuming that the least number of errors occurred.

**Solution.** The generating matrix of the code is

$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$



last element “bubble” back to the beginning), every permutation can be written as a composition of transpositions of adjacent elements.

Return to the case of  $\Sigma_3$ . Two elements,  $b, c$ , represent cycles which include all the three elements; each of them generates  $\{a, b, c\} = \mathbb{Z}_3$ . Besides those,  $d, e, f$  are composed of cycles of length 2 and 1; finally  $a$  is composed of three cycles of length one. There are no more possibilities. However, it is clear from the procedure that for more elements, there are very many possibilities.

In general, there are many ways of expressing a permutation as a composition of transpositions. However, for a given permutation, the parity of the number of transpositions is fixed and independent of the choice of particular transpositions. This can be seen from the number of inverses of a permutation, since each transposition changes the number of inverses by an odd number (see the discussion in subsection 2.2.2 on page 97).

It follows that there is a well-defined mapping

$$\text{sgn} : \Sigma_m \rightarrow \mathbb{Z}_2 = \{\pm 1\},$$

the *permutation parity*. This recovers the proposition crucial for building the determinants (see 2.2.1 and on):

**Theorem.** Every permutation of a finite set can be written as a composition of cycles. A cycle of length  $\ell$  can be expressed as a composition of  $\ell - 1$  transpositions. The parity of this cycle is  $(-1)^{\ell-1}$ .

The parity of the composition  $\sigma \circ \tau$  is equal to the product of the parities of the composed permutations  $\sigma$  and  $\tau$ .

The last proposition suggests that the mapping  $\text{sgn}$  transforms permutation composition  $\sigma \circ \tau$  to the product  $\text{sgn } \sigma \cdot \text{sgn } \tau$  in the commutative group  $\mathbb{Z}_2$ .

(SEMI)GROUP HOMOMORPHISMS

In general, a mapping  $f : G_1 \rightarrow G_2$  is a (semi)group homomorphism if and only if it respects the operation, i.e.

$$f(a \cdot b) = f(a) \cdot f(b).$$

In particular, the permutation parity is a homomorphism  $\text{sgn} : \Sigma_m \rightarrow \mathbb{Z}_2$ .

In a moment, we shall see that the group inversions and units are also preserved by homomorphisms. Before discussing the theory, let us look at more examples of groups.

**12.4.3. Symmetries of plane figures.** In the fifth part of chapter one, the connections between invertible 2-by-2 matrices and linear transformations in the plane are thoroughly considered.

A matrix in  $\text{Mat}_2(\mathbb{R})$  defines a linear mapping  $\mathbb{R}^2 \rightarrow \mathbb{R}^2$  that preserves standard distances if and only if its columns form an orthonormal basis of  $\mathbb{R}^2$  (which is a simple condition for the matrix entries, see subsection 1.5.7 on page 33). Combining the orthogonal linear mappings with translations, we arrive at the group of all Euclidean transformations of the plane.



The generating matrix is of the form  $G = \begin{pmatrix} P \\ \mathbb{I}_k \end{pmatrix}$ , where  $P =$

$\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$ . The check matrix is of the form  $(\mathbb{I}_{n-k} \ P)$ , i. e., in our case

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

Multiplying the received word by the check matrix, we get the syndrome (error) of the word:

$$H \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = (0 \ 1 \ 1 \ 1 \ 1).$$

The syndrome corresponding to the received word is 01111. Now, we find all words corresponding to this syndrome. This can be done by adding all codewords to the received word. There are four codewords, corresponding to the four possible messages. They are obtained by multiplying the messages (00, 01, 10, 11) by the generating matrix. Thus, we get the codewords

$$0000000, 1111101, 1010110, 0101011.$$

The space of words corresponding to a given syndrome is an affine space whose direction is the vector space of all codewords (see 12.5.8). Thus, we get the words

$$0010111, 1101010, 1000001, 0111100.$$

The least number of errors is equal to the least number of ones in the obtained words. In our case, this is the word 1000001, which contains only two ones and thus is the so-called leading representative of the class of words with syndrome 01111. The original message can be obtained by subtracting (or adding – this is equivalent in  $\mathbb{Z}_2$ ) the received word and the leading representative of the class with the given syndrome. In our case, we get

$$0010111 - 1000001 = 1010110.$$

Therefore, assuming the least number of errors, the sent word was 1010110, where the last two bits are the original message, i. e., 10.  $\square$

**12.H.7.** Consider the (7, 3)-code generated by the polynomial  $x^4 + x^3 + x + 1$ . Find its generating matrix and check matrix. Using the method of leading representatives, decode the received word 1110010.  $\circ$

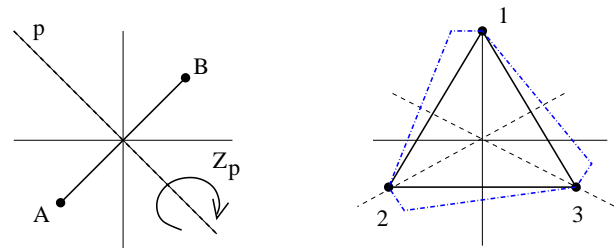
In fact, it is possible to prove that every mapping of the plane into itself which preserves distances is affine such a Euclidean transformation.<sup>9</sup>

As observed, the linear part of this mapping is orthogonal. Thus, all these mappings form a group of all orthogonal transformations (also called Euclidean transformations) in the plane. Moreover, it is shown that besides translations  $T_a$  by a vector  $a$ , these are only rotations  $R_\varphi$  around the origin by any angle  $\varphi$ , and reflection  $F_\ell$  with respect to any line that goes through the origin (also note that the central inversion is the same as the rotation by  $\pi$ ).



Now, general group concepts are illustrated on the problem of symmetries of plane figures. For example, consider tiles. First, consider them individually, in the form of a bounded diagram in the plane. Then consider with the condition of tiling in a band, and then in the entire plane.

As an example, consider a line segment and an equilateral triangle. It is of interest in how much these objects are symmetric; that is, with respect to which transformations (that preserve distances) they are invariant. In other words, we want the image of the figure to be identical to the original one (unless some significant points are labeled, for example the vertices of the triangle  $A, B, C$  or the endpoints of the line segment). It is clear that all symmetries of a fixed object form a group (usually with only one element, the identity).



In the case of the line segment, the situation is very simple – it is clear that the only non-trivial symmetries are rotation by  $\pi$  around the center of the segment, reflection  $F_H$  with respect to the axis of the segment, and reflection  $F_V$  with respect to the line itself. All these symmetries are self-inverse. Hence the group of symmetries has four elements. Its table looks as follows:

<sup>9</sup>If a mapping  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  preserves distances, then this must also hold for the mapped vectors of velocity, i.e. the Jacobi matrix  $DF(x, y)$  must be orthogonal at every point. Expanding this condition for the given mapping  $F = (f(x, y), g(x, y)) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  leads to a system of differential equations which has only affine solutions, since all second derivatives of  $F$  must be zero (and then, the proposition is an immediate consequence of Taylor's remainder theorem). Try to think out the details! The same procedure leads to the result for Euclidean spaces of arbitrary dimension. Note that the condition to be proved is independent of the choice of affine coordinates. Composing  $F$  with a linear mapping does not change the result. Hence, for a fixed point  $(x, y)$ , compose  $(DF)^{-1} \circ F$  and assume, without loss of generality, that  $DF(x, y)$  is the identity matrix. Differentiation of the equations then yields the desired proposition.



**12.H.8.** Consider the linear  $(7, 4)$ -code (i. e., the message has length 4) over  $\mathbb{Z}_2$  defined by the matrix

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Decode the received word 1010001 (i. e., find the sent message) assuming that the least number of errors occurred.

**Solution.** There are  $2^4 = 16$  possible messages. All codewords can be obtained by multiplying the possible messages (0000, 0001, ..., 1111) by the generating matrix of the code. Thus, we get:

$$\begin{array}{cccc} 0110001, & 1010010, & 1100100, & 0111000 \\ 1100011, & 1010101, & 0001001, & 1011100 \\ 1101010, & 0110110, & 0001110, & 1101101 \\ 1011011, & 0000111, & 0111111, & 0000000. \end{array}$$

Now, we construct the check matrix of the given code:

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

(we remove the block of the generating matrix that consists of the identity matrix and to the left of the remaining block, we write the identity matrix of fitting size). Now, multiplying the vector of the received word  $z^T = (1010001)$  by  $H$ , we get the syndrome  $s = Hz = (110)^T$ . One word with this syndrome is 1100000 (we fill the syndrome with zeros to the appropriate length). All words with syndrome 110 are obtained by adding this word to all codewords. Thus, we get the words

$$\begin{array}{cccc} 1000001, & 0110010, & 0000100, & 1011000, \\ 0000011, & 0110101, & 1101001, & 0111100, \\ 0001010, & 1010110, & 1101110, & 0001101, \\ 0111011, & 1100111, & 1011111, & 1100000 \end{array}$$

Out of these words with syndrome 110, only the word 0000100 contains a single one, so this is the leading representative of the class of words with syndrome 110. Subtracting the leading representative from the received word, we get the word that was sent, assuming the least number of bit flips (1 in this case), i. e., the word (101)0101, where the last four bits are the message, i. e. 0101.  $\square$

**12.H.9.** Consider the linear  $(7, 4)$ -code (i. e., the message has length 4) over  $\mathbb{Z}_2$  defined by the matrix

$$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\begin{array}{c|cccc} \cdot & R_0 & R_\pi & F_H & F_V \\ \hline R_0 & R_0 & R_\pi & F_H & F_V \\ R_\pi & R_\pi & R_0 & F_V & F_H \\ F_H & F_H & F_V & R_0 & R_\pi \\ F_V & F_V & F_H & R_\pi & R_0 \end{array}.$$

This group is commutative.

For the equilateral triangle, there are more symmetries: one can rotate by  $2\pi/3$  or one can mirror with respect to axes of the sides. In order to obtain the entire group, all compositions of these transformations must be added in. In 1.5.9 it is shown that the composition of two reflections is always a rotation. At the same time, it is clear that changing the order of composition of fixed two reflections leads to a rotation by the same angle but the other orientation. It follows that the reflections around two axes generate all the symmetries, of which there are six altogether. Placing the triangle as is shown in the diagram, the six transformations are given by the following matrices:

$$a = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \quad c = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix},$$

$$d = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad e = \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}, \quad f = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}.$$

A comparison of the table of the operation, with that of the permutation group  $\Sigma_3$ , shows that it is the same. For the sake of clarity, the vertices are labeled with numbers, so the corresponding permutations can be easily understood.

Similarly, there are groups of symmetries with  $k$  rotations and  $k$  reflections. It suffices to consider a regular  $k$ -gon. These groups are usually denoted  $D_k$  and are called the *dihedral groups* of order  $k$ . They are not commutative for  $k \geq 3$  ( $D_2$  is commutative). The name comes from the fact that  $D_2$  is the group of symmetries of the hydrogen molecule  $H_2$ , which contains two hydrogen atoms and can be imagined as a line segment.

Similarly, there are figures whose only symmetries are rotations, and hence the corresponding groups are commutative. They are denoted  $C_k$  and called *cyclic groups* of order  $k$ . For that, it suffices to consider a regular polygon whose sides are changed non-symmetrically, but in the same manner (see the extension of the triangle in the diagram). Note that the group  $C_2$  can be realized in two ways: either using the rotation by  $\pi$  or a single reflection.

As the first illustration of the power of abstraction, we prove the following theorem. A figure is said to have a *discrete group of symmetries* if and only if the set of images of an arbitrary point over all the symmetries is a discrete subset of the plane (i.e. each of its points has a neighbourhood where there is no other point of the set).

Note that every discrete group of symmetries of a bounded figure is necessarily finite.

**Theorem.** Let  $M$  be a bounded set in the plane  $\mathbb{R}^2$  with discrete group of symmetries  $G$ . Then,  $G$  is either trivial or one of the groups  $C_k, D_k$  for  $k > 1$ .

Decode the received word 1101001 (i. e., find the sent message) assuming that the least number of errors occurred.

**Solution.** Syndrome 101, leading representative 0001000, sent message (110)0001  $\square$

**12.H.10.** Consider the linear (7, 4)-code (i. e., the message has length 4) over  $\mathbb{Z}_2$  defined by the matrix

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Decode the received word 0000011 (i. e., find the sent message) assuming that the least number of errors occurred.

**Solution.** Syndrome 011, leading representative 0000100, sent message (000)0111.  $\square$

**12.H.11.** Consider the linear (7, 4)-code (i. e., the message has length 4) over  $\mathbb{Z}_2$  defined by the matrix

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Decode the received word 0001100 (i. e., find the sent message) assuming that the least number of errors occurred.

**Solution.** Syndrome 110, leading representative 0000010, sent message (000)1110.  $\square$

**12.H.12.** We want to transfer one of four possible messages with a binary code which should be able to correct all single errors. What is the minimum possible length of the codewords (all codewords have to be of the same length)? Why?



**Solution.** Let us denote the desired length as  $n$ . The minimum Hamming distance of any two codewords must be at least three. This means that if we take two different codewords and flip one bit in each, the resulting words must also be different (and also different from each codeword). There are  $n + 1$  words that can be obtained from a given one by flipping at most one bit (this includes the original word itself as well). Thus, we must have at least  $4(n + 1)$  possible words. On the other hand, there are  $2^n$  words of length  $n$ , which means that  $4(n + 1) \leq 2^n$ . This inequality is satisfied only for  $n \geq 5$ . Thus, the codewords must be at least 5 bits long. And indeed, there are four codewords of length 5 with minimum Hamming distance 3, for instance 00111, 01001, 10100, 11010.  $\square$

**PROOF.** If there were a set  $M$  with translation as one of its symmetries, then it could not be bounded. If  $M$  had, as one of its symmetries, a rotation by angle which is an irrational multiple of  $2\pi$ , then iterating this rotation would lead to a dense subset of images on the corresponding circle.<sup>10</sup> It follows that the group is not discrete.

If  $M$  had non-trivial rotations with different centers as symmetries, then again it could not be bounded. To see this, write the corresponding rotations in the complex plane as

$$R : z \mapsto (z - a)\zeta + a, \quad Q : z \mapsto z\eta$$

for complex units  $\zeta = e^{2\pi i/k}$ ,  $\eta = e^{2\pi i/\ell}$  and an arbitrary  $a \neq 0 \in \mathbb{C}$ . Then, it is immediate (a straightforward computation with complex numbers) that

$$Q \circ R \circ Q^{-1} \circ R^{-1} : z \mapsto z + a(-1 + \zeta + \eta - \zeta\eta),$$

which is a translation by a non-trivial vector unless the angle of one of the rotations is zero. It follows that  $M$  is not bounded.

The same holds for the case of a rotation and a reflection with respect to a line which does not go through the center of the rotation. Check this case yourself!

The only symmetries available are rotations with a common center and reflections with respect to lines which pass through this center. It remains to prove that the entire group is composed either only from rotations, or from the same number of rotations and reflections.

Recall that the composition of two different reflections yields a rotation whose angle doubles the angle enclosed by the corresponding axes (see 1.5.9). Therefore, composing a reflection with respect to a line  $p$  with a rotation by angle  $\varphi$  is again a reflection with respect to the line which is at angle  $\varphi/2$  from  $p$  (draw a diagram!).

The proof is almost complete. Observe that the subgroup of all rotations in the group of symmetries contains a rotation by the smallest positive angle  $\varphi_0$  (there are only finite many of them there). But then it is impossible to allow a rotation  $R_\varphi$  where  $\varphi$  is not a multiple of  $\varphi_0$  (for then  $\varphi \in (k\varphi_0, (k + 1)\varphi_0)$  for some  $k$  and the composition  $R_{-k\varphi_0} \circ R_\varphi$  would have an smaller angle than  $\varphi_0$ ). This subgroup coincides with one of  $C_\ell$ . Next, adding one reflection produces exactly one different reflection for each nontrivial element in  $C_\ell$ , as seen above.  $\square$

**12.4.4. Symmetries of plane tilings.** There is more complicated behaviour in the case of plane figures in bands or in the entire plane (for example, symmetries of various tilings).



<sup>10</sup>The argument is subtle but straightforward: if there was an interval of length  $\varepsilon$  on the circle not hit by an orbit of a point under the rotation, all points in the orbit would have to be at distance at least  $\varepsilon$ . Thus there can be only finitely many of them and this contradicts the irrationality of the angle.

**12.H.13.** We want to transfer 4-bit messages with a binary code which should be able to correct all single and double errors. What is the minimum possible length of the codewords (all codewords have to be of the same length)? Why?



**Solution.** We proceed similarly as in the above exercise. If the code is to correct double errors as well, then the minimum Hamming distance of any two codewords must be at least three. This means that if we take two different codewords and flip up to two bits in each, the resulting words must also be different. Denoting by  $n$  the length of the words, we get the inequality

$$2^4 \left( 1 + n + \binom{n}{2} \right) \leq 2^n.$$

The least value of  $n$  for which it is satisfied is  $n = 12$ , so the codewords must be at least 12 bits long.  $\square$

### I. Extension of the stereographic projection

Let us try to extend the definition of the stereographic projection so that a circle would be parametrized by points of  $\mathbb{P}_1(\mathbb{R})$ . Let us look at the corresponding mapping  $\mathbb{P}_1(\mathbb{R}) \rightarrow \mathbb{P}_2(\mathbb{R}^2)$ . The points in projective extensions will be defined in the so-called *homogeneous coordinates*, which are given up to a multiple. For instance, the points in  $\mathbb{P}_2(\mathbb{R})$  will be  $(x : y : z)$ .

A circle in the plane  $z = 1$  is given as the intersection of the cone of directions defined by  $x^2 + y^2 - z^2 = 0$  with this plane. The inversion of the stereographic projection (i. e., our parametrization of the circle) can be described as:

$$(t : 1) \mapsto \left( \frac{2t}{1+t^2} : \frac{t^2-1}{t^2+1} : 1 \right) = (2t : t^2 - 1 : t^2 + 1).$$

For  $t \neq 0$ , we have  $(t : 1) = (2t^2 : 2t)$ , and the original stereographic projection (i. e., the inverse of the above mapping) can be written linearly as

$$(x : y : z) \mapsto (y + z : x),$$

which extends our parametrization to the improper point  $(0 : 1) \mapsto (0 : 1 : 1)$ . Then, the mapping of  $\mathbb{P}_1(\mathbb{R})$  onto the circle has the “linear” form

$$\mathbb{P}_1(\mathbb{R}) \ni (x : y) \mapsto (2x : x - y : x + y) \in \mathbb{P}_2(\mathbb{R}).$$

Now, let us look at how simple it is to calculate the formula for the stereographic projection in the projective extensions directly (see 4.4.1): We include  $\mathbb{P}_1(\mathbb{R})$  as points with homogeneous coordinates  $(t : 0 : 1)$ , and among the linear combinations of point  $(0 : 1 : -1)$  (i. e., the pole from which we project) and  $(x : y : z)$  (a general point of the circle), we must find the one whose coordinates are  $(u : 0 : v)$ . The only possibility is the point  $(x : 0 : z + y)$ , which is our previous formula.

Consider first the set containing the points that lie between two fixed parallel lines. Suppose that this band is covered with disjoint images of a bounded subset  $M$  by some translation. Of course, this translation is a symmetry of the chosen tiling of the band. So the group of symmetries is necessarily infinite.

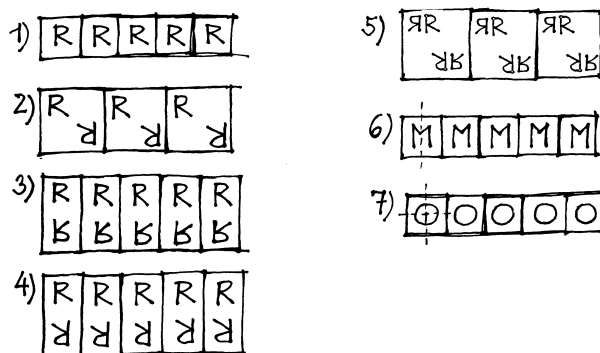
Such a set allows for no other rotation symmetries than  $R_\pi$ , and the only possible reflections are either horizontal with respect to the axis of the band, or vertical with respect to any line which is perpendicular to the boundary lines. In addition, there are translations given by a vector which is parallel to the axis of the band.

A not-too-complicated discussion leads to description of all discrete groups of symmetries for these bands. Such a group is generated by some of the following symmetries: translation  $T$ , shifted reflection  $G$  (i.e. composition of horizontal reflection and translation), vertical reflection  $V$ , horizontal reflection  $H$  and rotation  $R$  by  $\pi$ .

**Theorem.** Every discrete group of symmetries of a band in the plane is isomorphic to one of the groups generated by the following symmetries:

- (1) a single translation  $T$ ,
- (2) a single shifted reflection  $G$ ,
- (3) a single translation  $T$  and a vertical reflection  $V$ ,
- (4) a single translation  $T$  and the rotation  $R$ ,
- (5) a single shifted reflection  $G$  and the rotation  $R$ ,
- (6) a single translation  $T$  and the horizontal reflection  $H$ ,
- (7) a single translation  $T$ , the horizontal reflection  $H$  and a vertical reflection  $V$ .

The proof is not presented here. The following diagram shows examples of schematic patterns with corresponding symmetries:



It is even more complicated with symmetries of tilings which cover the entire plane. There is insufficient space here to consider further details. It can be shown that there are 17 such groups of symmetries, known as the two-dimensional crystallographic groups.

A similar complete discussion is known even for three-dimensional discrete groups of symmetries. The rich theory was created namely in the 19th century in connection with studying symmetries of crystals and molecules of chemical elements.

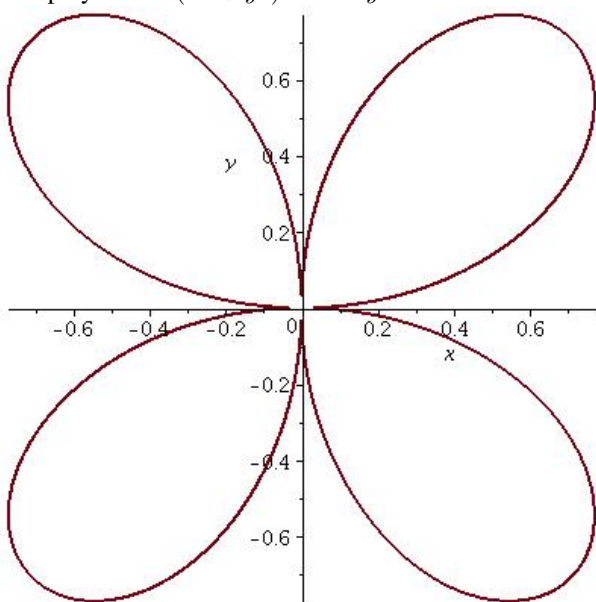
**J. Elliptic curves**

A *singular point* of a hypersurface in  $P^n$ , defined by a homogeneous polynomial

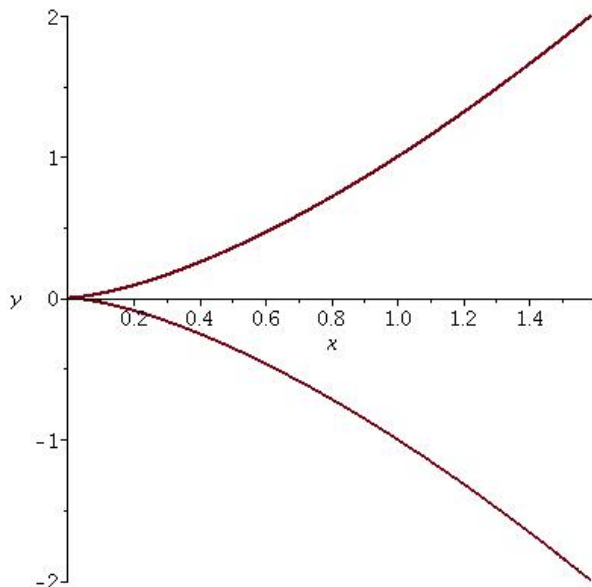
$$F(x_0, x_1, \dots, x_n) = 0,$$

is such a point which satisfies  $\frac{\partial F}{\partial x_i} = 0$  for  $i = 1, \dots, n$ .

From the geometric point of view, “something weird” happens at the point. In the case of a curve in the projective space  $P^2(\mathbb{R})$ , the condition that the partial derivatives must be zero means that there is no tangent line to the curve at the given point. This means that the curve has the so-called *cusp* there or it intersects itself. A “nice” singularity can be seen in the “quatrefoil”, i. e., the variety given by the zero points of the polynomial  $(x^2 + y^2)^3 - 4x^2y^2 \in \mathbb{R}^2$ :



The cusp can be found on the curve in  $\mathbb{R}^2$  given by  $x^3 - y^2 = 0$ .



**12.4.5. Group homomorphisms.** Recall that a mapping  $f : G \rightarrow H$  from a group  $G$  to a group  $H$  is called a group homomorphism if and only if it respects the operation, i.e. for all  $a, b \in G$  that



$$f(a \cdot b) = f(a) \cdot f(b).$$

Note that the operation on the left-hand side is the operation in  $G$ , before  $f$  is applied, while the operation on the right-hand side is the operation in  $H$ , after  $f$  is applied.

The following properties of homomorphisms follow easily from the definition:

**Proposition.** Every group homomorphism  $f : G \rightarrow H$  satisfies:

- (1) the identity of  $G$  is mapped to the identity of  $H$ ,
- (2) the inverse of an element of  $G$  is mapped to the inverse of its image, i.e.

$$f(a^{-1}) = f(a)^{-1},$$

- (3) the image of a subgroup  $K \subseteq G$  is a subgroup  $f(K) \subseteq H$ ,
- (4) the preimage  $f^{-1}(K) \subseteq G$  of a subgroup  $K \subseteq H$  is again a subgroup,
- (5) if  $f$  is also a bijection, then the inverse mapping  $f^{-1}$  is also a homomorphism,
- (6)  $f$  is injective if and only if  $f^{-1}(e_H) = \{e_G\}$ .

**PROOF.** (3), (2), and (1). If  $K \subseteq G$  is a subgroup, then for each  $y = f(a), z = f(b)$  in  $H, a, b \in K, y \cdot z = f(a \cdot b)$  also lies in the image. In particular,  $f(b) = f(e \cdot b) = f(e) \cdot f(b)$  and similarly  $f(b) = f(b) \cdot f(e)$ . Thus  $f(e)$  is the unique unit in the image  $f(K)$ . Finally,  $f(e) = f(a \cdot a^{-1}) = f(a) \cdot f(a^{-1})$ , so that  $f(a^{-1})$  is the right inverse of  $f(a)$ . Similarly, it is a left inverse, too, and we have proved the first three claims.

Proceed similarly in the case of preimages: if  $a, b \in G$  satisfy  $f(a), f(b) \in K \subseteq H$ , then also  $f(a \cdot b) \in K$ .

Suppose there exists an inverse mapping  $g = f^{-1}$ . Fix arbitrarily  $y = f(a), z = f(b) \in H$ . Then,  $f(a \cdot b) = y \cdot z = f(a) \cdot f(b)$ , which is equivalent to the expression  $g(y) \cdot g(z) = a \cdot b = g(y \cdot z)$ . Thus the inverse mapping is also a homomorphism.

If  $f(a) = f(b)$ , then  $f(a \cdot b^{-1}) = e_H$ . Therefore, if the only element that is mapped to  $e_H$  is  $e_G$ , then  $a \cdot b^{-1} = e_G$ , i.e.  $a = b$ . The other implication is trivial.  $\square$

The subgroup  $f^{-1}(e_H)$  (the preimage of the identity in  $H$ ) is called the *kernel* of the homomorphism  $f$  and is denoted  $\ker f$ . A bijective group homomorphism is called a *group isomorphism*.

It follows directly from the above ideas that a homomorphism  $f : G \rightarrow H$  with a trivial kernel is an isomorphism onto the image  $f(G)$ .

An *elliptic curve*  $\mathcal{C}$  is the set of points in  $\mathbb{K}^2$ , where  $\mathbb{K}$  is a given field, which satisfy an equation of the form

$$y^2 = x^3 + ax + b,$$

where  $a, b \in \mathbb{K}$ . In addition, we require that there are no singularities, which means, over the field of real numbers, that

$$\Delta = -16(4a^3 + 27b^2) \neq 0.$$

This expression  $\Delta$  is called the discriminant of the equation. Note that the right-hand side contains a cubic polynomial without the quadratic term. This form of the equation is called the *Weierstrass equation* of an elliptic curve.

**12.J.1.** Prove that the curve  $y^2 = x^3 + ax + b$  in  $\mathbb{R}^2$  has a singularity if and only if  $4a^3 - 27b^2 = 0$ .

**Solution.** The equation of the curve in homogeneous coordinates (see 4.4.1) is  $F(x, y, z) = 0$ , where

$$(1) \quad F(x, y, z) = y^2z - x^3 - axz^2 - bz^3.$$

We have

$$\begin{aligned} \frac{\partial F}{\partial x} &= -3x^2 - az^2, \\ \frac{\partial F}{\partial y} &= 2yz, \\ \frac{\partial F}{\partial z} &= y^2 - 2axz - 3bz^2. \end{aligned}$$

Let  $[x, y, z]$  be a singular point of the given curve. If  $z = 0$ , then since the partial derivatives of  $F$  with respect to  $x$  and  $z$  must be zero, we get  $x = 0$  and  $y = 0$ , respectively. However, this is “out”, because the point  $[0, 0, 0]$  does not lie in the considered projective space  $P^2(\mathbb{R})$ . Thus, the singular points has  $z \neq 0$ , so that  $\frac{\partial F}{\partial y} = 0$  implies  $y = 0$ . Denoting  $\gamma = \frac{x}{z}$ , then  $-3x^2 - az^2 = 0$  implies  $3\gamma^2 = -a$ , and  $y^2 - 2axz - 3bz^2 = 0$  implies  $2a\gamma = -3b$ . We can see that the equality  $a = 0$  implies that  $b = 0$ , i. e., the equality  $4a^3 = 27b^2$  is satisfied trivially. If  $a \neq 0$ , then we can express  $\gamma$  from the two obtained equations. From the first one, we have  $\gamma = -\frac{3b}{2a}$ , and from the second one,  $\gamma^2 = -\frac{a}{3}$ . Altogether,

$$\gamma^2 = -\frac{a}{3} = \frac{9b^2}{4a^2} \implies 4a^3 + 27b^2 = 0.$$

Thus, we have proved one of the implications. On the other hand, if  $4a^3 + 27b^2 = 0$ , then defining  $\gamma = -\frac{3b}{2a}$ , we can see that the point  $[\gamma, 0, 1]$  satisfies the equation of the elliptic curve:

$$\begin{aligned} \gamma^2 + a\gamma + b &= \left(-\frac{3b}{2a}\right) \left(-\frac{a}{3}\right) + a \left(-\frac{3b}{2a}\right) + b = \\ &= \frac{b}{2} - \frac{3b}{2} = 0. \end{aligned}$$

Thanks to the choice of  $\gamma$ , all the three partial derivatives of  $F$  at the point  $[\gamma, 0, 1]$  are zero.  $\square$

In order to define a group operation on the points of an elliptic curve, it is useful to consider the curve in the projective extension of the plane (see 4.4.1), and we define a point

**12.4.6. Examples.** The additive group  $\mathbb{Z}_k$  of residue classes modulo  $k$  is isomorphic to the group of  $k$ -th roots of unity, and also to the group of rotations by integer multiples of  $2\pi/k$ . Draw a diagram, calculation with the complex units  $e^{2\pi i/k}$  is very efficient.



The mapping  $\exp : \mathbb{R} \rightarrow \mathbb{R}_+$  is an isomorphism of the additive group of the real numbers onto the multiplicative group of the positive real numbers.

This isomorphism extends naturally to a homomorphism  $\exp : \mathbb{C} \rightarrow \mathbb{C} \setminus \{0\}$  of the additive group of the complex numbers onto the multiplicative group of the non-zero complex numbers. However, this homomorphism has a non-trivial kernel. The restriction of  $\exp$  to purely imaginary numbers (which is a subgroup isomorphic to  $\mathbb{R}$ ) is a homomorphism

$$it \mapsto e^{it} = \cos t + i \sin t.$$

This means that the numbers  $2k\pi i$ ,  $k \in \mathbb{Z}$ , lie in the kernel. It can be shown that nothing else is in the kernel. If  $e^{s+it} = e^s \cdot e^{it} = 1$  for real numbers  $s$  and  $t$ , then  $e^s = 1$ , i.e.  $s = 0$ , and then  $t = 2k\pi$  for an integer  $k$ .

The determinant of a matrix is a mapping which assigns, to each square matrix of scalars in  $\mathbb{K}$ , a scalar in  $\mathbb{K}$  (the cases  $\mathbb{K} = \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$  have already been worked with). The Cauchy theorem about the determinant of the product of square matrices

$$\det(A \cdot B) = (\det A) \cdot (\det B)$$

can be also seen as the fact that for the group  $G = GL(n, \mathbb{K})$  of invertible matrices, the mapping  $\det : G \rightarrow \mathbb{K} \setminus \{0\}$  is a group homomorphism.

**12.4.7. Group product.** Given any two groups, a more complicated group can be constructed using the following construction:

#### GROUP PRODUCT

For any groups  $G, H$  the *group product*  $G \times H$  is defined as follows: The underlying set is the Cartesian product  $G \times H$  and the operation is defined componentwise. That is,

$$(a, x) \cdot (b, y) = (a \cdot b, x \cdot y),$$

where the left-hand operation is the one to be defined, while the right-hand operations are respectively those in  $G$  and  $H$ .

The projections onto the components  $G$  and  $H$  in the product:

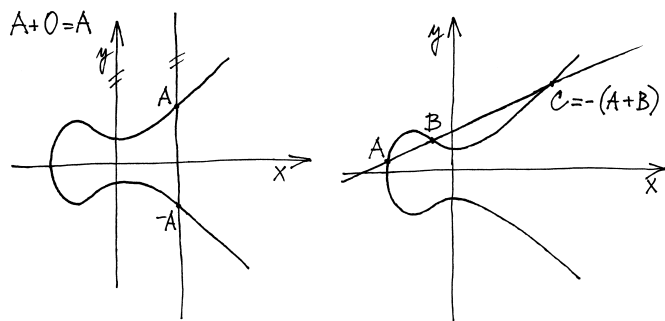
$$p_G : G \times H \ni (a, b) \mapsto a \in G, \quad p_H : G \times H \ni (a, b) \mapsto b$$

are surjective homomorphisms, whose kernels are

$$\begin{aligned} \ker p_G &= \{(e_G, b); b \in H\} \simeq H, \\ \ker p_H &= \{(a, e_H); a \in G\} \simeq G. \end{aligned}$$

The group  $\mathbb{Z}_6$  is isomorphic to the product  $\mathbb{Z}_2 \times \mathbb{Z}_3$ . This can be seen easily in the multiplicative realization of the groups  $\mathbb{Z}_k$  as the complex  $k$ -th roots of unity.  $\mathbb{Z}_6$  consists of the points of the unit circle that form the vertices of a regular hexagon. Then,  $\mathbb{Z}_2$  corresponds to  $\pm 1$ , while  $\mathbb{Z}_3$  corresponds to the equilateral triangle, one of whose vertices is

$O \in \mathcal{C}$  as the direction  $(0, 1)$  (which is the point  $[0, 1, 0]$  in the homogeneous coordinates). Then, the addition of two points  $A, B \in \mathcal{C}$  is geometrically defined as the point  $-C$ , where  $C$  is the third intersection point of the line  $AB$  with the elliptic curve. If  $A = B$ , then the result is given by the other intersection point with the tangent line of the elliptic curve that goes through  $A$ .



**12.J.2.** Prove that the above definition correctly defines an operation on the points of an elliptic curve.

**Solution.** The intersections of the line with the elliptic curve are obtained as the roots of a cubic equation. If it has two real roots, corresponding to the points  $A$  and  $B$ , then it must have a third real root as well, i. e., the line  $AB$  must have another intersection point with the curve. In the case of a tangent line, the point  $A$  corresponds to a double root, so there also exists another intersection point. As for improper points (the last homogeneous coordinate is zero; they correspond to directions in the plane), the only improper point that belongs to the curve given by the equation (1) is the point  $O = [0, 1, 0]$ . Addition with the point  $O$  means looking for a second intersection of the elliptic curve (besides the point  $A$  itself) and the line which goes through  $A$  and is parallel to the  $y$ -axis. The improper line  $z = 0$  has triple intersection point  $O$  with the curve, i. e.,  $O + O = O$ .  $\square$

**Remark.** Thus, the operation is well-defined. Moreover, it follows directly from the definition that it is commutative. It even follows from the above that  $O$  is a neutral element of the operation. However, the proof of associativity is far from trivial.

**12.J.3.** Define the above operation algebraically.

**Solution.** For any point  $A \in \mathcal{C}$ , we define  $A + O = O + A = A$ .

For a point  $A \in \mathcal{C}$ ,  $A = (\alpha, \beta, 1)$ , we clearly have  $B \in \mathcal{C}$ , and we define  $A + B = 0$ , i. e.,  $A = -B$ .

For a point  $A \neq -B$ ,  $A = [\alpha, \beta, 1]$  and a point  $B \in \mathcal{C}$ ,  $B = [\gamma, \delta, 1]$ , we set

$$k = \begin{cases} \frac{\beta - \delta}{\alpha - \gamma} & \text{for } A \neq B, \\ [5pt] \frac{3\alpha^2 + a}{2\beta} & \text{for } A = B, \end{cases}$$

$$\sigma = k^2 - \alpha - \gamma,$$

$$\tau = -\beta + k(\alpha - \sigma).$$

the number 1. If each point is identified with the rotation that maps 1 to that point, then the composition of such rotations is always commutative. Composing a rotation from  $\mathbb{Z}_2$  with a rotation from  $\mathbb{Z}_3$  yields exactly all rotations from  $\mathbb{Z}_6$ . Draw a diagram! This leads to the following isomorphism (using additive notation, as is common for the residue classes):

$$\begin{aligned} [0]_6 &\mapsto ([0]_2, [0]_3), \\ [1]_6 &\mapsto ([1]_2, [2]_3), \\ [2]_6 &\mapsto ([0]_2, [1]_3), \\ [3]_6 &\mapsto ([1]_2, [0]_3), \\ [4]_6 &\mapsto ([0]_2, [2]_3), \\ [5]_6 &\mapsto ([1]_2, [1]_3). \end{aligned}$$

Similar constructions are available for finite commutative groups in complete generality.

**12.4.8. Commutative groups.** Any element  $a$  of a group



$G$  is contained in the minimal subgroup  $\{\dots, a^{-2}, a^{-1}, e, a, a^2, a^3, \dots\}$ , which contains it. It is clear that this subgroup is commutative. If  $G$  is finite, then it must once happen that  $a^k = e$ . The least positive integer  $k$  with this property is called the *order of the element  $a$*  in  $G$ . A *cyclic group*  $G$  is one which is generated by one of its elements  $a$  in the above manner. If the order  $k$  of the generator is finite, then it results in one of the groups  $C_k$ , known from the discussion of symmetries of plane figures.

It follows directly from the definition that every cyclic group is isomorphic either to the group of integers  $\mathbb{Z}$  (if it is infinite) or to one of the groups of residue classes  $\mathbb{Z}_k$  (if it is finite). These simple building stones are sufficient to create all finite commutative groups.

**Theorem.** Every finite commutative group  $G$  is isomorphic to a product of some cyclic groups  $C_k$ . The orders of the components  $C_k$  are always powers of the prime divisors of the number of elements  $n = |G|$ . This product decomposition is unique, up to order.

If  $n = p_1^{k_1} \dots p_r^{k_r}$  is the prime factorization of  $n$ , then the group  $C_n$  is isomorphic to the product

$$C_n = C_{p_1^{k_1}} \times \dots \times C_{p_r^{k_r}}.$$

**INCOMPLETE PROOF.** We are going to prove only the second claim now, and we return to the first claim later, see 12.4.12, 12.4.13.

For a simpler case, suppose  $n = pq$  with  $p$  coprime to  $q$ . Fix a generator  $a$  of the group  $C_n$ , a generator  $b$  of  $C_p$ , and a generator  $c$  of  $C_q$ . Define the mapping  $f : C_n \rightarrow C_p \times C_q$  by

$$f(a^k) = (b^k, c^k).$$

Since  $a^k \cdot a^\ell = a^{k+\ell}$  and similarly for  $b$  and  $c$ , it follows that

$$f(a^k \cdot a^\ell) = (b^{k+\ell}, c^{k+\ell}) = (b^k, c^k) \cdot (b^\ell, c^\ell),$$

so the mapping  $f$  is a homomorphism. If the image is the identity, then  $k$  must be a multiple of  $p$  as well as  $q$ . Since

Then, we define  $A + B = [\gamma, \tau, 1]$ . We leave it for the reader to verify that this is indeed the operation that we have defined geometrically.  $\square$

**K. Gröbner bases**

**12.K.1.** Is the basis  $g_1 = x^2, g_2 = xy + y^2$  a Gröbner basis for the lexicographic ordering  $x > y$ ? If not, find one.

**Solution.** Clearly, the leading monomials are  $LT(g_1) = x^2, LT(g_2) = xy$ , so the S-polynomial is equal to

$$S(g_1, g_2) = yg_1 - xg_2 = -xy^2.$$

By theorem 12.6.12,  $g_1, g_2$  is a Gröbner basis if and only if the remainder in the multivariate division of this S-polynomial by the basis polynomials is zero. Performing this division (see 12.6.6), we obtain

$$S(g_1, g_2) = yg_1 - xg_2 + yg_2 - y^3.$$

The remainder  $y^3$  shows that  $g_1, g_2$  do not form a Gröbner basis. By 12.6.13, in order to get one, we must add the remainder polynomial  $g_3 = y^3$  to  $g_1, g_2$ . Now, we calculate that

$$S(g_1, g_3) = y^3g_1 - x^2g_3 = 0$$

and

$$S(g_2, g_3) = y^2g_2 - xg_3 = y^4 = yg_3.$$

Hence it follows by theorem 12.6.12 that  $g_1, g_2, g_3$  is already a Gröbner basis.  $\square$

**12.K.2.** Is the basis  $g_1 = xy - 2y, g_2 = y^2 - x^2$  a Gröbner basis for the lexicographic ordering  $y > x$ ? If not, find one.

**Solution.** Since  $LT(g_1) = xy$  and  $LT(g_2) = y^2$ , the corresponding S-polynomials is  $S(g_1, g_2) = yg_1 - xg_2 = x^3 - 2y^2 = -2g_2 + x^3 - 2x^2$ . The leading term  $x^3$  is a multiple of neither  $xy$  nor  $y^2$ , which means that  $g_1, g_2$  do not form a Gröbner basis. We can obtain one by adding the polynomial  $g_3 = x^3 - 2x^2$ . Then, we have

$$S(g_1, g_3) = x^2g_1 - yg_3 = 0$$

and

$$S(g_2, g_3) = x^3g_2 - y^2g_3 = 2y^2x^2 - x^5 = (4y + 2xy)g_1 - (x^2 + 2x + 4)g_3 + 8g_2. \quad \square$$

**12.K.3.** Eliminate variables in the ideal

$$I = \langle x^2 + y^2 + z^2 - 1, x^2 + y^2 + z^2 - 2x, 2x - y - z \rangle.$$

**Solution.** The variable elimination is obtained by finding a Gröbner basis with respect to the lexicographic monomial ordering. Let us denote the generating polynomials of  $I$  as  $g_1, g_2, g_3$ , respectively. The reduction  $g_2 = g_1 + 1 - 2x$  yields the reduced polynomial  $f_1 = 2x - 1$ . Now, we use this polynomial to reduce  $g_3 = f_1 + 1 - y - z$  to  $f_2 = y + z - 1$ . Now, we reduce  $g_1$ , dividing it by  $f_1$  and  $f_2$ , which leads to

$$g_1 = \left(\frac{1}{2}x + \frac{1}{4}\right)f_1 + y^2 + z^2 - 1$$

and

$$y^2 + z^2 - 1 = (y - z + 1)f_2 + 2z^2 - 2z + \frac{1}{4}.$$

$p$  and  $q$  are coprime,  $k$  is a multiple of  $n$ , so  $f$  is injective. Moreover, the group  $C_n$  has the same number of elements as  $C_p \times C_q$ , so  $f$  is an isomorphism.

Finally, the proposition about the decomposition of cyclic groups of order  $k$  into smaller cyclic groups follows by induction on the number of different primes  $p_i$  in the factorization of  $n$ .  $\square$

Notice,  $C_{p^2}$  is never isomorphic to the product  $C_p \times C_p$ . While  $C_{p^2}$  is generated by an element of order  $p^2$ , the highest order of an element in  $C_p \times C_p$  is only  $p$ .

Since every finite commutative group is isomorphic to a product of cyclic groups, it is possible, for a given number of elements, to enumerate all commutative groups of that order up to isomorphism. For instance, there are only two groups of order 12:

$$C_{12} = C_4 \times C_3, \quad C_2 \times C_2 \times C_3 = C_2 \times C_6.$$

Notice similarly that if all elements (except the identity) of a finite commutative group  $G$  have order 2, then  $G$  has the form  $(C_2)^n$  for an integer  $n$ . In particular, such a group  $G$  has  $2^n$  elements. If the decomposition of  $G$  into cyclic groups contains a group  $C_p, p > 2$ , then the group contains elements of higher order.

**12.4.9. Subgroups and cosets.** Selecting any subgroup  $H$



of a group  $G$ , gives further information about the structure of the whole group. A binary relation  $\sim_H$  on  $G$  can be defined as follows:  $a \sim_H b$  if and only if  $b^{-1} \cdot a \in H$ . This relation expresses when two elements of  $G$  are “the same” up to multiplication by an element of  $H$  from the right. It is easily verified that this relation is an equivalence:

Clearly,  $a^{-1} \cdot a = e \in H$ , so it is reflexive. If  $b^{-1} \cdot a = h \in H$ , then  $a^{-1} \cdot b = (b^{-1} \cdot a)^{-1} = h^{-1} \in H$ , so it is symmetric as well. Finally, if  $c^{-1} \cdot b \in H$  and  $b^{-1} \cdot a \in H$ , then  $c^{-1} \cdot a = c^{-1} \cdot b \cdot b^{-1} \cdot a \in H$ , so it is transitive, too.

It follows that  $G$  partitions into the *left cosets* of mutually equivalent elements, with respect to the subgroup  $H$ . The coset corresponding to an element  $a$  is denoted  $a \cdot H$ , and

$$a \cdot H = \{a \cdot h; h \in H\},$$

since an element  $b$  is equivalent to  $a$  if and only if it can be expressed this way.

The corresponding partition of  $G$  (i.e. the set of all left cosets) with respect to  $H$  is denoted  $G/H$ .

Similarly, right cosets  $H \cdot a$  can be defined. The corresponding equivalence relation is given by  $a \sim b$  if and only if  $a \cdot b^{-1} \in H$ . Hence,

$$H \setminus G = \{H \cdot a; a \in G\}.$$

**Proposition.** Let  $G$  be a group and  $H$  a subgroup of  $G$ . Then:

(1) The left cosets with respect to  $H$  coincide with the right cosets with respect to  $H$  if and only if for each  $a \in G$ ,

Hence,  $f_3 = 8z^2 - 8z + 1$ . We can see that we could do with polynomial reduction and we did not have to add any other polynomials. The basis of  $I$  with eliminated variables is  $I = \langle 2x - 1, y + z - 1, 8z^2 - 8z + 1 \rangle$ .  $\square$

**12.K.4.** Solve the following system of polynomial equations:

$$\begin{aligned} x^2y - z^3 &= 0, \\ 2xy - 4z &= 1, \\ z - y^2 &= 0, \\ x^3 - 4yz &= 0. \end{aligned}$$

**Solution.** Using appropriate software, we can find out that the corresponding ideal

$$\langle x^2y - z^3, 2xy - 4z - 1, z - y^2, x^3 - 4yz \rangle$$

has Gröbner basis  $\langle 1 \rangle$  with respect to the lexicographic monomial ordering, which means that the system has no solution.  $\square$

**12.K.5.** Find the Gröbner basis of the variety in  $\mathbb{R}^3$  defined parametrically as



$$\begin{aligned} x &= 3u + 3uv^2 - u^3, \\ y &= 3v + 3u^2v - v^3, \\ z &= 3u^2 - 3v^2. \end{aligned}$$

This is the so-called Enneper surface, and it is depicted in the picture on page 1119.

**Solution.** Applying the elimination procedure (e. g. in the MAPLE system, using `gbasis` with `plex` ordering), we obtain the corresponding implicit representation, i. e., an equation with a single polynomial of degree nine:

$$\begin{aligned} &-59049z - 104976z^2 - 6561y^2 - 72900z^3 - 18954y^2z - \\ &-23328z^4 + 32805z^2x^2 + 14580z^3x^2 + 3645z^4x^2 - 1296y^4z - \\ &-16767y^2z^2 - 6156y^2z^3 - 783y^2z^4 + 39366zx^2 + 19683x^2z - \\ &-1296y^4 - 2430z^5 + 432z^6 + 108z^7 + 486z^5x^2 - 432y^4z^2 + \\ &+ 54y^2z^5 + 27z^6x^2 - 48y^4z^3 + 15y^2z^6 - 64y^6 - z^9. \end{aligned}$$

$\square$  As we illustrate on the following simple exercise, the Gröbner basis can be used for solving integer optimization problems.

**12.K.6.** What is the minimum number of banknotes that are necessary to pay 77700 CZK? Solve this problem for three scenarios: First, assume that the banknotes at disposal are of values 100 CZK, 200 CZK, 500 CZK, 1000 CZK. Then, assume that there are also banknotes of value 2000 CZK. Finally, assume that there are no banknotes of value 2000 CZK, but there are banknotes of value 5000 CZK.

**Solution.** Let us denote the respective banknotes by variables  $s, d, p, t, D, P$ . The banknotes to be used will be represented as a polynomial in these variables so that the exponent of each variable determines the number of the corresponding banknotes. For instance, if we decide to use only the 100 CZK banknotes, then the polynomial will be  $q = s^{777}$ . If we pay with ten 1000 CZK banknotes, ten 500 CZK banknotes,

$$h \in H$$

$$a \cdot h \cdot a^{-1} \in H.$$

(2) Each coset (left or right) has the same cardinality as the subgroup  $H$ .

**PROOF.** Both properties are direct consequences of the definition. In the first case, for any  $a \in G, h \in H$ , an element  $h' \in H$  is required so that  $h \cdot a = a \cdot h'$ . This occurs if and only if  $a^{-1} \cdot h \cdot a = h' \in H$ .

In the second case, if  $a \cdot h = a \cdot h'$ , then multiplication by  $a^{-1}$  from the left yields  $h = h'$ .  $\square$

As an immediate corollary of the above statement, there are the following extremely useful results:

**12.4.10. Theorem.** Let  $G$  be a finite group with  $n$  elements and  $H$  a subgroup of  $G$ . Then:

(1) the cardinality  $n = |G|$  is the product of the cardinality of  $H$  and the cardinality of  $G/H$ , i.e.

$$|G| = |G/H| \cdot |H|,$$

(2) the integer  $|H|$  divides  $n$ ,

(3) if  $a \in G$  is of order  $k$ , then  $k$  divides  $n$ ,

(4) for each  $a \in G, a^n = e$ ,

(5) if  $n$  is prime, then  $G$  is isomorphic to the cyclic group  $\mathbb{Z}_n$ .

The second proposition is called *Lagrange's theorem*. The fourth proposition is called *Fermat's little theorem*. Special cases are discussed in the previous chapter on number theory.

**PROOF.** Each left coset has exactly  $|H|$  elements. However, different cosets are disjoint. Hence the first proposition follows.

The second proposition is a direct corollary of the first one.

Each element  $a \in G$  generates the cyclic subgroup  $\{a, a^2, \dots, a^k = e\}$ , and the order of this subgroup is exactly the order of  $a$ . Therefore, the order of  $a$  must divide the number of elements of  $G$ .

Since the order  $k$  of any element  $a$  divides  $n$  and  $a^k = e$ , then also  $a^n = (a^k)^s = e$  for any integer  $s$ .

If  $n > 1$ , then there exists an element  $a \in G$  that is different from  $e$ . Its order  $k$  is an integer greater than one and it divides  $n$ . Therefore,  $k$  must be equal to  $n$ . This means that all the elements of  $G$  are of the form  $a^\ell$  for  $\ell = 1, \dots, n$ .  $\square$

**12.4.11. Normal subgroups and quotient groups.** A subgroup  $H$  which satisfies  $a \cdot h \cdot a^{-1} \in H$  for all  $a \in G, h \in H$ , is called a *normal subgroup*.



For normal subgroups, the operation on  $G/H$  can be defined by

$$(a \cdot H) \cdot (b \cdot H) = (a \cdot b) \cdot H.$$

Choosing other representatives  $a \cdot h, b \cdot h'$  leads to the same result:

$$(a \cdot h \cdot b \cdot h') \cdot H = ((a \cdot b) \cdot (h^{-1} \cdot h \cdot b) \cdot h') \cdot H = (a \cdot b) \cdot H.$$



and the remaining amount with 100 CZK banknotes, then the polynomial will be  $q = t^{10}p^{10}s^{627}$ . In the former case, the number of banknotes will be 777. In the latter case, it will be  $10 + 10 + 627 = 647$ .

If we have only the banknotes  $s, d, p, t$ , then the ideal that describes the relation of the individual banknotes is

$$I_1 = \langle s^2 - d, s^5 - p, s^{10} - t \rangle.$$

In order to minimize the number of used banknotes, we compute the Gröbner basis with respect to the graded reverse lexicographic ordering (we want to eliminate the small banknotes):

$$G_1 = (p^2 - t, s^2 - d, d^3 - sp, sd^2 - p).$$

Now, we take any polynomial that represents a given choice of banknotes. Reducing this polynomial with respect to the basis  $G_1$ , we get a polynomial whose degree is minimal for our monomial ordering, at it is easy to show that it is the polynomial corresponding to the optimal choice. For instance, take  $q = s^{777}$ . Reduction with respect to  $G_1$  yields  $t^{77}pd$ . This means that the optimal choice is seventy-seven 1000 CZK banknotes, one 500 CZK banknote, and one 200 CZK banknote. Altogether, it is 79 banknotes.

In the second scenario, when we also have the banknote  $D$ , the ideal is  $I_2 = \langle s^2 - d, s^5 - p, s^{10} - t, s^{20} - D \rangle$  and its Gröbner basis is

$$G_2 = (t^2 - D, p^2 - t, s^2 - d, d^3 - sp, sd^2 - p).$$

Reduction of  $q = s^{777}$  with respect to  $G_2$  gives  $D^{38}tpd$ , so this time we pay with 41 banknotes. In the third scenario, we have  $I_3 = \langle s^2 - d, s^5 - p, s^{10} - t, s^{50} - P \rangle$  and

$$G_3 = (t^5 - P, p^2 - t, s^2 - d, d^3 - sp, sd^2 - p),$$

and the reduction is equal to  $P^{15}t^2pd$ . In this case, we need only 19 banknotes.

Of course, this simple problem can be solved quickly with common sense. However, the presented method of Gröbner bases gives a universal algorithm which can be automatically used for higher amounts and other, more complicated cases.  $\square$

Gröbner bases have applications in robotics as well. In particular, it is in inversion kinematics, where one must find how to set individual joints of a robot so that it could reach a given position. This problem often leads to a system of non-linear equations which can be solved by finding a Gröbner basis, as in the following problem.

**12.K.7.** Consider a simple robot, as shown in the picture, which consists of three straight parts of length 1 which are connected with independent joints that enable arbitrary angles  $\alpha, \beta, \gamma$ . We want this robot to grasp, from above, an object which lies on the ground in distance  $x$ . What values should the angles  $\alpha, \beta, \gamma$  be set to? Draw the configuration of the robot for  $x = 1, 1.5a\sqrt{3}$ .

**Solution.** Consider a natural coordinate system, where the initial end of the robotic hand lies in the origin, and the

Moreover, cosets can be written as  $H \cdot a \cdot H$ , and the equation  $(H \cdot a) \cdot (b \cdot H) = H \cdot (a \cdot b) \cdot H$  is straightforward. On the other hand, the definition of the product on the cosets fails if  $H$  is not normal.

Clearly, this new operation on  $G/H$  satisfies all group axioms: the identity is the group  $H$  itself (formally it is the coset  $e \cdot H$  that corresponds to the identity  $e$  of  $G$ ), the inverse of  $a \cdot H$  is  $a^{-1} \cdot H$ , and the associativity is clear from the definition. This is called the *quotient group*  $G/H$  of  $G$  by the normal subgroup  $H$ .

Of course, in commutative groups, every subgroup is normal. The subset

$$n\mathbb{Z} = \{na; a \in \mathbb{Z}\} \subseteq \mathbb{Z}$$

is a subgroup of the integers, and the corresponding quotient group is the (additive) group  $\mathbb{Z}_n$  of residue classes.

It is clear from the definition that the kernel of every homomorphism is a normal subgroup. On the other hand, if a subgroup  $H \subseteq G$  is normal, then the mapping

$$p : G \rightarrow G/H, \quad a \mapsto a \cdot H$$

is a surjective homomorphism, whose kernel is  $H$ . It can be seen directly from the definition of the operation on  $G/H$  that  $p$  is a homomorphism, and it is clearly surjective. It follows that normal subgroups are precisely the kernels of homomorphisms.

Moreover, for any group homomorphism  $f : G \rightarrow K$  with kernel  $H = \ker f$ , there is a well-defined homomorphism

$$\tilde{f} : G/H \rightarrow K, \quad \tilde{f}(a \cdot H) = f(a),$$

which is injective. If  $H$  is any normal subgroup in  $G$  contained in  $\ker f$ , the latter homomorphism is still well defined, but not necessarily injective.

There is a seemingly paradoxical example of a group homomorphism  $\mathbb{C}^* \rightarrow \mathbb{C}^*$ , defined on the non-zero complex numbers by  $z \mapsto z^k$ , where  $k$  is a fixed positive integer. Clearly, this is a surjective homomorphism, and its kernel is the set of  $k$ -th roots of unity, i.e. the cyclic subgroup  $\mathbb{Z}_k$ . Reasoning as above, there is an isomorphism

$$\tilde{f} : \mathbb{C}^*/\mathbb{Z}_k \rightarrow \mathbb{C}^*$$

for any positive integer  $k$ . This example illustrates that in the case of infinite groups, the calculations with cardinalities are not so intuitive as in the case of finite groups and theorem 12.4.10.

**12.4.12. Exact sequences.** A normal subgroup  $H$  of a group  $G$  yields the *short exact sequence of groups*

$$e \rightarrow H \rightarrow G \rightarrow G/H \rightarrow e,$$

where the arrows respectively correspond to the only homomorphism of the trivial group  $\{e\}$  into the group  $H$ , the inclusion  $\iota$  of the subgroup  $H \subseteq G$ , the projection  $\nu$  onto the quotient group  $G/H$ , and the only homomorphism of the group  $G/H$  onto the trivial group  $\{e\}$ . In each case, the image of one arrow is precisely the kernel of the following one. This is the definition of *exactness* of a sequence of homomorphisms.

ground corresponds to the  $x$ -axis. It follows from elementary trigonometry that the total  $x$ -range of the robot at angles  $\alpha, \beta, \gamma$  is equal to

$$x = \sin \alpha + \sin(\alpha + \beta) + \sin(\alpha + \beta + \gamma).$$

Similarly, the range of the robot in the vertical direction is

$$y = \cos \alpha + \cos(\alpha + \beta) + \cos(\alpha + \beta + \gamma).$$

The condition of grasping the object from above is clearly equivalent to the condition  $\alpha + \beta + \gamma = \pi$ , so the statement of the problem leads to the system

$$\begin{aligned} \sin \alpha + \sin(\alpha + \beta) &= x, \\ \cos \alpha + \cos(\alpha + \beta) - 1 &= 0. \end{aligned}$$

In order to transform this system to a system of polynomial equations, we introduce new variables  $s_1 = \sin \alpha$ ,  $c_1 = \cos \alpha$ ,  $s_2 = \sin \beta$ ,  $c_2 = \cos \beta$ , which of course satisfy  $s_1^2 + c_1^2 = 1$  and  $s_2^2 + c_2^2 = 1$ . Using basic trigonometric equalities for sums in arguments, we obtain the following, equivalent system of polynomial equations:

$$\begin{aligned} s_1 + s_1 c_2 + c_1 s_2 - x &= 0, \\ c_1 + c_1 c_2 - s_1 s_2 - 1 &= 0, \\ s_1^2 + c_1^2 - 1 &= 0, \\ s_2^2 + c_2^2 - 1 &= 0. \end{aligned}$$

The Gröbner basis of the corresponding ideal can be found in a while using appropriate software. For the graded reverse lexicographic ordering  $s_1 > c_1 > s_2 > c_2$ , we get the basis

$$\begin{aligned} (2c_2 + 1 - x^2, 2c_1(1 + x^2) - 2s_2x - 1 - x^2, 2s_1(1 + x^2) + \\ + 2s_2 - x - x^3, 4s_2^2 - 3 - 2x^2 + x^4), \end{aligned}$$

and hence it is easy to calculate the values of the variables in dependence on  $x$ . For example, we can immediately see that  $c_2 = \frac{x^2-1}{2}$ , i. e.,  $\beta = \arccos(\frac{x^2-1}{2})$ . In particular, it is clear that the problem has no solution for  $|x| > \sqrt{3}$ . Specifically, for  $|x| < \sqrt{3}$ , there are 2 solutions, and for  $|x| = \sqrt{3}$ , there is one solution ( $\alpha = \frac{\pi}{3}, \beta = 0, \gamma = \frac{2\pi}{3}$  for positive  $x$ ;  $\alpha = -\frac{\pi}{3}, \beta = 0, \gamma = \frac{4\pi}{3}$  for negative  $x$ ). For  $x = 1$ , we get the solution  $\alpha = 0, \beta = \frac{\pi}{2}, \gamma = \frac{\pi}{2}$  and the degenerated solution  $\alpha = \frac{\pi}{2}, \beta = -\frac{\pi}{2}, \gamma = \pi$ . The case  $x = -1$  is similar. It is good to realize that for  $|x| < 1$ , one of the solutions will always correspond to a configuration of the robot where some parts will intersect. For these values of  $x$ , there will be only one realizable configuration.

If there exists a homomorphism  $\sigma : G/H \rightarrow G$  such that  $\nu \circ \sigma = \text{id}_{G/H}$ , it is said that the exact sequence *splits*.

**Lemma.** *Every split short exact sequence of commutative groups defines an isomorphism  $G \rightarrow H \times G/H$ .*

**PROOF.** Define a mapping  $f : H \times G/H \rightarrow G$  by

$$f(a, b) = a \cdot \sigma(b).$$

Since the groups are commutative,  $f$  is a homomorphism:

$$f(aa', bb') = aa' \sigma(b) \sigma(b') = (a \sigma(b)) (a' \sigma(b')).$$

If  $f(a, b) = e$ , then  $\sigma(b) = a^{-1} \in H$ , i.e.  $b = \nu(\sigma(b))$  is the identity in  $G/H$ . However, its image is then  $\sigma(b) = e$ , so  $a = e$ . Since the left and right cosets of commutative groups coincide, the mapping  $f$  is surjective. Hence  $f$  is an isomorphism.  $\square$

Now, the main idea of the proof of theorem 12.4.8 can be indicated. If it is known that every short exact sequence created by choosing cyclic subgroups  $H$  of a finite commutative group  $G$  splits, then it is easy to proceed with the proof by induction. If  $G$  is a group of order  $n$  which is not cyclic, then an element  $a$  of order  $p$ ,  $p < n$ , can be selected. The cyclic subgroup  $H$  generated by  $a$  can be found as well as the splitting of the corresponding short exact sequence. This expresses the group  $G$  as the product of the selected cyclic subgroup  $H$  and the group  $G/H$  of order  $n/p$ .

The main technical point of the proof is the verification that in each finite commutative group, there are elements of order  $p^r$  with appropriate powers of the primes  $p$  and that the short exact sequences for these groups really split.

**12.4.13. Return to finite Abelian groups.** Below is a brief exposition of the complete proof of the classification theorem, broken into several steps.



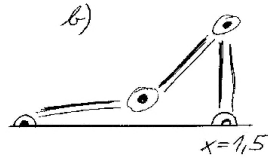
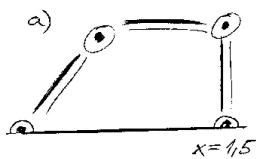
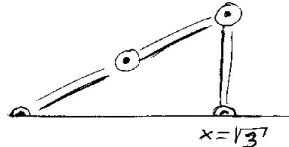
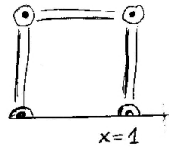
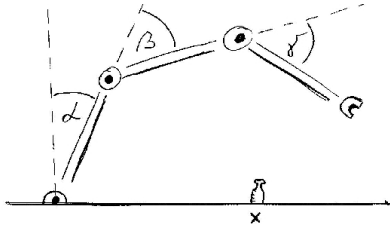
The following lemma suggests that cyclic subgroups with prime orders are required.

**Lemma (Claim 1).** *Let  $G$  be a finite Abelian group with  $n$  elements. If  $p$  is a prime which divides  $n$ , then there is an element  $g \in G$  of order  $p$ .*<sup>11</sup>

**PROOF.** The claim is obvious if  $n$  is prime, i.e.  $G = \mathbb{Z}_p$  (as proved above). If  $n$  is not prime, proceed by induction on  $n$ . Clearly  $G$  must have a proper subgroup  $H$  if  $n$  is not prime,  $|H| = m < n$ . Either  $p|m$  or  $p|(n/m)$ . In the former case, the claim follows from the induction hypothesis directly.

Otherwise assume  $p|(n/m)$  (remind  $n/m$  is the order of  $G/H$ ). Then there is an element  $g \in G$  such that the order of  $g \cdot H$  in the quotient group  $G/H$  is  $p$ . Thus  $g^p \in H$ , and therefore the order of  $g$  in  $G$  divides  $p|H|$ . Since  $p$  is a prime, the order of  $g$  is  $\ell p$  for some integer  $\ell$ . Hence the element  $g^\ell$  has the required order  $p$ .  $\square$

<sup>11</sup>This is a special version of the more general result valid for all finite groups, called the Cauchy theorem. The formulation remains the same, with the word Abelian omitted.



Gröbner bases can also be used in software engineering when looking for loop invariants, which are needed for verification of algorithms, as in the following problem.

**12.K.8.** Verify the correctness of the algorithm for the product of two integers  $a, b$ .

```

(x, y, z) := (a, b, 0);
while not (y = 0) do
  if y mod 2 = 0
  then (x, y, z) := (2*x, y/2, z)
  else (x, y, z) := (2*x, (y-1)/2, x+z)
  end if
end while
return z
    
```

**Solution.** Let  $X, Y, Z$  denote the initial values of the variables  $x, y, z$ , respectively. Then, by definition, a polynomial  $p$  is an invariant of the loop if and only if we have  $p(x, y, z, X, Y, Z) = 0$  after each iteration. Such a polynomial can be found using Gröbner basis as follows:

Let  $f_1, f_2$  denote the assignments of the then- and else-branches, respectively, i. e.,

$$f_1(x, y, z) = (2x, \frac{1}{2}y, z) \text{ and } f_2(x, y, z) = (2x, \frac{y-1}{2}, x+z).$$

For  $n$  iterations of the first one, we immediately obtain the explicit formula  $f_1^n(x, y, z) = (2^n x, \frac{1}{2^n} y, z)$ . In order to transform this iterated function to a polynomial mapping, we introduce new variables  $u := 2^n, v := \frac{1}{2^n}$ . Then,  $f_1^n$  is given by the polynomial function

$$F_1: x \mapsto ux \quad y \mapsto vy \quad z \mapsto z,$$

For any prime number  $p$ ,  $G$  is called a  $p$ -group if each of its elements has order  $p^k$  for some power  $k$ . Claim 1 has an obvious corollary:

**Lemma (Claim 2).** A finite Abelian group  $G$  is a  $p$ -group if and only if its number of elements  $n$  is a power of  $p$ .

**PROOF.** One implication follows straight from the Lagrange's theorem since all proper divisors of a power of a prime  $p$  are just smaller powers of  $p$ .

On the other hand, if  $n$  is not a power of prime, it has another prime divisor  $q$  and so there is an element of order  $q$  by Claim 1.  $\square$

Now it can be shown that a given finite Abelian group  $G$  can always be decomposed into a product of  $p$ -groups.

**Lemma (Claim 3).** If  $G$  is a finite Abelian group then it is isomorphic to a product of  $p$ -groups. This decomposition is unique up to order.

**PROOF.** Consider a prime  $p$  dividing  $n = |G|$ . Define  $G_p$  to be the subgroup of all elements whose orders are powers of  $p$ , while  $G'_p$  is the subgroup of all elements whose orders are not divisible by  $p$  (check yourself that subgroups are obtained in this way). By the above Claim 1, the subgroup  $G_p$  is not trivial.

Next, consider an element  $g$  of order  $qp^\ell$  with  $q$  not divisible by  $p$ . Then  $g^{p^\ell}$  has order  $q$ , so this element belongs to  $G'_p$ , while  $g^q \in G_p$ . The Bezout equality guarantees there are integers  $r$  and  $s$  such that  $rp^\ell + sq = 1$ . Hence

$$g = g^{rp^\ell} \cdot g^{sq}$$

is a decomposition of  $g$  into product of elements in  $G_p$  and  $G'_p$ . This verifies  $G \simeq G_p \times G'_p$  and  $G_p$  is a  $p$ -group.

This process can be repeated for the subgroup  $G'_p$  and the remaining primes in the decomposition in order to complete the proof.

The uniqueness claim is obvious.  $\square$

It remains to consider the  $p$ -groups only. The next claim shows that the  $p$ -groups which are not cyclic must have more than one subgroup of order  $p$ .

**Lemma (Claim 4).** If a finite Abelian  $p$ -group  $G$  has just one subgroup  $H$  with  $|H| = p$ , then it is cyclic.

**PROOF.** The case  $p = n = |G|$  is obvious. Proceed by induction on  $n$ . Assume  $H$  is the only subgroup of order  $p$  and consider  $\sigma : G \rightarrow G, \sigma(g) = g^p$  and write  $K = \ker(\sigma)$ . Then  $H \subset K$  and since  $p$  is prime, all elements in  $K$  have order  $p$ . For any  $g \in K$ , the cyclic group generated by  $g$  has order  $p$  and so coincides with  $H$  and consequently  $H = K$ . If  $G \neq K$ , then  $\sigma(G)$  is a non-trivial subgroup in  $G$  which must be isomorphic to  $G/K$ . By Claims 1 and 2, there is a subgroup in  $\sigma(G)$  of order  $p$ . This yields a subgroup in  $G$  and by assumption it is again  $H$ . Finally, apply the induction hypothesis on the group  $\sigma(G) \simeq G/H$ , which has to be cyclic. Choosing a generator  $g \cdot H$  of the latter group, even in  $G$  the

where the new variables satisfy  $uv = 1$ . Clearly, the invariant polynomial must lie in the ideal

$$I_1 = \langle ux - X, vy - Y, z - Z, uv - 1 \rangle.$$

In order to find such polynomial, it suffices to eliminate the variables  $u$  and  $v$ , which can be done just with the Gröbner basis with respect to the graded reverse lexicographic ordering with  $u > v > x > y > z$ . This basis is equal to

$$(xy - XY, z - Z, x - vX, y - uY).$$

Hence  $F_1(xy - XY) = xy - XY$  and  $F_1(z - Z) = z - Z$ , and all other polynomials are invariant with respect to any number  $n$  of applications of  $f_1$  and are given by a polynomial in (polynomials)  $xy - XY$  and  $z - Z$ .

Now, we proceed similarly for  $f_2$ . For  $n$  iterations, we derive the formula

$$f_2^n(x, y, z) = (2^n x, \frac{1}{2^n}(y + 1) - 1, (2^n - 1)x + z),$$

and introducing the variables  $u$  and  $v$ , we get an equivalent polynomial function

$$F_2: \quad x \mapsto ux \quad y \mapsto v(y + 1) - 1 \quad z \mapsto (u - 1)x + z.$$

The invariant polynomial for  $F_2$  can be obtained similarly as above, thanks to the Gröbner basis of the corresponding ideal. However, we are interested in those polynomials which are invariant for both  $F_1$  and  $F_2$ . Clearly, these must lie in the ideal

$$I_2 = \langle F_2(xy - XY), F_2(z - Z), uv - 1 \rangle.$$

Substituting for  $F_2$ , we obtain

$$I_2 = \langle u xv(y + 1) - ux - XY, (u - 1)x + z - Z, uv - 1 \rangle$$

and with the Gröbner basis of this ideal, we eliminate the variables  $u$  and  $v$  and thus find the polynomial  $xy - XY + z - Z$ , which is invariant for both  $F_1$  and  $F_2$ , so it is an invariant of the given cycle. Since at the beginning we have  $X = a, Y = b, Z = 0$ , we can see that it holds in every step of the algorithm that  $xy - ab + z = 0$ . Since the loop terminates only if  $y = 0$ , we get that indeed  $z = ab$ .  $\square$

Now, we present several exercises where we use Gröbner bases to solve various polynomial systems. The primary goal will not be to find the Gröbner basis, but rather to solve the given system.

**12.K.9.** Using Gröbner basis, solve the polynomial system

$$\begin{aligned} x^3 - 2xy &= 0, \\ x^2y + x - 2y^2 &= 0. \end{aligned}$$

**Solution.** Let us denote  $f_1 := x^3 - 2xy, f_2 := x^2y + x - 2y^2$ . The basis  $(f_1, f_2)$  is not a Gröbner basis since, e. g.,  $LM(yf_1 - xf_2) = x^2 \notin (x^3, x^2y) = (LM(f_1), LM(f_2))$ . Thus, we have to add the polynomial

$$yf_1 - xf_2 = -x^2.$$

cyclic subgroup generated by  $g$  must have a subgroup of order  $p$  (again Claim 1). The uniqueness assumption ensures that this is again the subgroup  $H$ .

Clearly,  $|G/H| = |G|/p$  is the smallest exponent with  $g^{|G|/p} \in H$ . At the same time the latter power is not equal to the unit since  $H \subset \langle g \rangle$ . Consequently, the order of  $g$  in  $G$  is bigger than  $|G|/p$  and so the group  $G$  is cyclic.  $\square$

Finally, a splitting condition for the  $p$ -groups is proved, which provides the property discussed in the end of the previous paragraph on the exact sequences. This completes the entire proof of the classification theorem.

**Lemma (Claim 5).** *Let  $G$  be a finite Abelian  $p$ -group and let  $C$  be a cyclic subgroup of maximal order in  $G$ . Then  $G = C \times L$  for some subgroup  $L$ .*

**PROOF.** If  $G$  is cyclic, set  $G = C$  and of course  $G = C \times L$  with  $L = \{e\}$ . Proceed by induction on  $n = |G|$ . Assume  $G$  is not cyclic. Then it contains more than one cyclic subgroup of order  $p$ . Of course the subgroup  $C$  is one such subgroup. Choose  $H$  to be another subgroup of order  $p$  which is not a subgroup in  $C$ . Since  $p$  is prime, the intersection of  $H$  and  $C$  is trivial. Consequently the quotient group  $(C \times H)/H \subset G/H$  is isomorphic to  $C$ .

Now consider the induction step. The order of the cyclic subgroup  $(C \times H)/H$  in  $G/H$  must be maximal, since the orders of the elements  $g \cdot H$  in the quotient group are divisors of the orders of the generators  $g$  in the group  $G$ . By the induction hypothesis,

$$G/H = (C \times H)/H \times K$$

for some subgroup  $K \subset G/H$ . Clearly the preimage of  $K$  under the quotient projection is a group  $L$  satisfying  $H \subset L \subset G$ . Now, the latter identification of  $G/H$  with the product implies

$$G = (C \cdot H) \cdot L = C \cdot (H \cdot L) = C \cdot L.$$

At the same time,  $L \cap (C \cdot H) = H$  and so  $L \cap C = \{e\}$ . So  $G = C \times L$ .  $\square$

The proof is complete up to the uniqueness claim. It is known already that the decomposition into  $p$ -groups is unique. Assume that a  $p$ -group  $G$  decomposes into two products of cyclic groups  $H_1 \times \dots \times H_k$  and  $H'_1 \times \dots \times H'_\ell$  with non-increasing orders of  $H_i$  or  $H'_j$ . Then the orders of  $H_1$  and  $H'_1$  coincide, since these are the maximal orders in  $G$ . By induction all the orders coincide and the work is complete.

The classification theorem is a special case of a more general result on finitely generated Abelian groups. In additive notation, if  $g_1, \dots, g_t$  are generators of the entire  $G$ , then all elements of  $G$  are of the form  $a_1g_1 + \dots + a_tg_t$  with integer coefficients  $a_i$ . The general theorem provides a severe restriction for possible relations between such combinations. In fact it says that all finitely generated Abelian groups are products of cyclic groups, hence  $G = \mathbb{Z}^\ell \oplus \mathbb{Z}_{p_1} \times \dots \times \mathbb{Z}_{p_k}$ . This means there is always a finite number of completely independent generators of  $G$  and each of them generates a cyclic subgroup in

The resulting basis can be reduced using division of the polynomials  $f_1, f_2$  by  $x^2$ . This leads to the basis

$$(xy, x - 2y^2, x^2).$$

Now, we can divide the first polynomial by the second one, with remainder  $2y^3$ , and the third one by the second one, with remainder  $4y^4$ . Thus, we get the basis

$$(x - 2y^2, y^3),$$

and this is a Gröbner basis: by the naive algorithm (see 12.6.13), it suffices to verify that the polynomial

$$S(x - 2y^2, y^3) = y^3(x - 2y^2) - xy^3 = -2y^5$$

gives zero remainder with respect to the basis  $(x - 2y^2, y^3)$ , even for any polynomial ordering. Clearly, the solution of the system is the point  $(0, 0)$ .  $\square$

**12.K.10.** Consider the following system of polynomial equations:

$$\begin{aligned} x^2yz^2 + x^2y^2 + yz - xyz^2 - z^2 &= 0, \\ x^2y + z &= 0, \\ xyz + z + 1 &= 0. \end{aligned}$$

Sort the monomials of the polynomials with respect to the lexicographic ordering with  $x > y > z$ , then divide the first polynomial by the second one and the third one, and use the result in order to solve the system in the real numbers.

**Solution.**

$$\begin{aligned} x^2y^2 + x^2yz^2 - xyz^2 + yz - z^2 &= (y + z^2)(x^2y + z) - \\ &\quad - y(xyz + z + 1) - z^3 + z. \end{aligned}$$

Hence  $z = 0, \pm 1$ . Then, e. g.,

$$\begin{aligned} 0 &= z(x^2y + z) - x(xyz + z + 1) = \\ &= z^2 - zx - x. \end{aligned}$$

Hence,  $x = \frac{z^2}{z+1}$ , and we get from the third equation that  $y = -\frac{(1+z)^2}{z^3}$ . This is satisfied by the sole point  $(\frac{1}{2}, -4, 1)$ .  $\square$

**12.K.11.** Using Gröbner basis, solve the polynomial system

$$\begin{aligned} x^2 + y + z &= 1, \\ x + y^2 + z &= 1, \\ x + y + z^2 &= 1. \end{aligned}$$

**Solution.** Let us denote  $f_1 := x + y + z^2 - 1$ . The division of  $x + y^2 + z - 1$  by  $f_1$  gives

$$f_2 = y^2 - y - z^2 + z.$$

The division of  $x^2 + y + z - 1$  by  $f_1$  yields  $y^2 + 2yz^2 - y + z^4 - 2z^2 + z$ , and further division by  $f_2$  produces the remainder.

$$f_3 = 2yz^2 + z^4 - z^2.$$

$G$ . (Compare this to the description of finite dimensional vector spaces via their basis, as discussed in chapter 2.)

**12.4.14. Group actions.** Groups can be considered as sets



of transformations of a fixed set. All the transformations are invertible, and the set of transformations must be closed under composition. The idea is to work with a fixed group whose elements are represented as mappings on a fixed set, but the mappings corresponding to different elements of the group need not be different. For instance, the rotations around the origin over all possible angles correspond to the group of real numbers. On the other hand, the rotation by  $2\pi$  is the identity as a mapping.

Formally, this situation can be described as follows:

#### GROUP ACTIONS

A *left action of a group  $G$  on a set  $S$*  is a homomorphism of the group  $G$  to the subgroup of invertible mappings in the monoid  $S^S$  of all mappings  $S \rightarrow S$ . Such a homomorphism can be viewed as a mapping  $\varphi : G \times S \rightarrow S$  which satisfies

$$\varphi(a \cdot b, x) = \varphi(a, \varphi(b, x)),$$

hence the name “left action”. Often, the notation  $a \cdot x$  is used to refer to the result of an  $a \in G$  applied to a point  $x \in S$  (although this is a different dot than the operation inside the group). Then, the definition property can be expressed as

$$(a \cdot b) \cdot x = a \cdot (b \cdot x).$$

The image of a point  $x \in S$  in the action of the entire group  $G$  is called the *orbit*  $S_x$  of  $x$ , i.e.

$$S_x = \{y = \varphi(a, x); a \in G\}.$$

For each point  $x \in S$ , define the *isotropy subgroup*  $G_x \subseteq G$  of the action  $\varphi$  (also called the *stabilizer subgroup*):

$$G_x = \{a \in G; \varphi(a, x) = x\}.$$

If for every two points  $x, y \in S$ , there is an  $a \in G$  such that  $\varphi(a, x) = y$ , then the action  $\varphi$  is said to be *transitive*.

Choosing any two points  $x, y \in S$  and a  $g \in G$  which maps  $x$  to  $y = g \cdot x$ , then the set  $\{ghg^{-1}; h \in G_x\}$  is clearly the isotropy subgroup  $G_y$ . In addition, the mapping  $h \mapsto ghg^{-1}$  is a group homomorphism  $G_x \rightarrow G_y$ .

In the case of transitive actions, the entire space forms a single orbit and all isotropy subgroups have the same cardinality.

As an example of a transitive action of a finite group, consider the apparent action of the symmetric group of a fixed set  $X$  on  $X$  itself. The natural action of all invertible linear transformations on the non-zero elements of a vector space  $V$  is also transitive. However, if the entire space  $V$  is selected, then the zero vector forms its own orbit.

The mentioned example of the action of the additive group of real numbers that acts as rotations around a fixed center  $O$  in the plane is not transitive. The orbit of each point  $M$  is the circle which is centered at  $O$  and goes through  $M$ .

However,  $(f_1, f_2, f_3)$  is not a Gröbner basis yet. That will be constructed by the choice  $g_1 := f_1, g_2 := f_2$  and replacing  $f_3$  with the  $S$ -polynomial

$$2z^2 f_2 - y f_3 = -yz^4 - yz^2 - 2z^4 + 2z^3.$$

Then, the division by the polynomial  $f_3$  leads to

$$\begin{aligned} g_4 &= z^6 - 4z^4 + 4z^3 - z^2 = \\ &= z^2(z-1)^2(z^2 + 2z - 1). \end{aligned}$$

Now,  $(g_1, g_2, g_3)$  is a Gröbner basis, so we can solve the system by elimination. We get from  $g_4 = 0$  that  $z = 0, 1, -1 \pm \sqrt{2}$ . Substituting this to  $g_2 = 0$  and  $g_1 = 0$  gives the solutions  $(1, 0, 0), (0, 1, 0), (0, 0, 1), (-1 + \sqrt{2}, -1 + \sqrt{2}, -1 + \sqrt{2}), (-1 - \sqrt{2}, -1 - \sqrt{2}, -1 - \sqrt{2})$ .  $\square$

**12.K.12.** Solve the following system of polynomial equations in  $\mathbb{R}$ :



$$\begin{aligned} x^2 - 2xz - 4, \\ x^2 y^2 z + yz^3, \\ 2xy^2 - 3z^3. \end{aligned}$$

**Solution.** The basis suitable for variable elimination is the Gröbner basis for the lexicographic monomial ordering with  $x > y > z$ . Using Maple, we can find the basis

$$\begin{aligned} 144z^5 + 35z^7 + 12z^9, \\ 23z^6 + 12z^8 + 44yz^4, \\ yz^3 + 3z^5 + 4zy^2, 9z^4 + 4y^3, \\ -8y^2 - 6z^4 + 3xz^3, 2xy^2 - 3z^3, \\ x^2 - 2xz - 4. \end{aligned}$$

Since the discriminant of the first polynomial of the basis (divided by  $z^5$ ) satisfies  $35^2 - 4 \cdot 144 \cdot 7 < 0$ , we must have  $z = 0$ . Substituting this into the other polynomials, we immediately obtain  $y = 0, x = \pm 2$ .  $\square$

**12.K.13.** Solve the following system of polynomial equations in  $\mathbb{R}$ :

$$\begin{aligned} xy + yz - 1, \\ yz + zw - 1, \\ zw + wx - 1, \\ wx + xy - 1. \end{aligned}$$

**Solution.** In this case, it is a good idea to take the graded lexicographic ordering with  $w > x > y > z$ . Using the algorithm 12.6.13 or appropriate software, we find the corresponding Gröbner basis

$$(x - z, w - y, 2yz - 1).$$

Thus, the system is satisfied by exactly the points  $(\frac{1}{2t}, t, \frac{1}{2t}, t)$  for an arbitrary  $t \in \mathbb{R}$  except zero.  $\square$

A typical example of a transitive action of a group  $G$  is the natural action on the set  $G/H$  of left cosets for any subgroup  $H$ . It is defined by

$$g \cdot (aH) = (ga)H.$$

This is the form of every transitive group action. For any transitive action  $G \times S \rightarrow S$  and a fixed point  $x \in G, S$  can be identified with the set  $G/G_x$  of left cosets by  $gG_x \mapsto g \cdot x$ . Clearly, this mapping is surjective, and the images  $g \cdot x = h \cdot x$  coincide if and only if  $h^{-1}g \in G_x$ , which is equivalent to  $gG_x = hG_x$ . Finally, note that this identification transforms the original action of  $G$  on  $S$  just to the mentioned action of  $G$  on  $G/G_x$ .

**12.4.15. Theorem.** Let an action of a finite group  $G$  on a finite set  $S$  be given. Then:

(1) for each point  $x \in S$ ,

$$|G| = |G_x| \cdot |S_x|,$$

(2) (Burnside's lemma) if  $N$  denotes the number of orbits, then

$$|G| = \frac{1}{N} \sum_{g \in G} |S^g|,$$

where  $S^g = \{x \in S; g \cdot x = x\}$  denotes the set of fixed points of the action corresponding to an element  $g$ .

**PROOF.** Consider any point  $x \in S$  and its isotropy subgroup  $G_x \subseteq G$ . The same argument as the one at the end of the previous paragraph for transitive group actions can be applied to each action of the group  $G$ . This gives the mapping  $G/G_x \rightarrow S_x, g \cdot G_x \mapsto g \cdot x$ . If  $g \cdot x = h \cdot x$ , then clearly  $g^{-1}h \in G_x$ , so this mapping is injective. Clearly, it is also surjective, which means that the cardinalities of the finite sets must satisfy  $|G/G_x| = |S_x|$ . The first proposition follows, because  $|G| = |G/G_x| \cdot |G_x|$ .

The second proposition is proved by counting the cardinality of the set of fixed points of individual group elements in two different ways:

$$F = \{(x, g) \in S \times G; g \cdot x = x\} \subseteq S \times G.$$

Since these are finite sets, the elements of the Cartesian product  $S \times G$  can be considered as entries of a matrix (columns are indexed by elements of  $S$ , rows are indexed by elements of  $G$ ). Summing this matrix up, either by rows or by columns, yields

$$|F| = \sum_{g \in G} |S^g| = \sum_{x \in S} |G_x|.$$

For the sake of clarity, choose one representative for each orbit (denote them  $x_1, \dots, x_N$ ), and recall that the cardinalities of the isotropy groups for points that lie in the same orbit are the same. Using the proved statement (1), it is obtained easily that

$$|F| = \sum_{g \in G} |S^g| = \sum_{i=1}^N \sum_{x \in S_{x_i}} |G_x| = \sum_{i=1}^N |S_{x_i}| |G_{x_i}| = N \cdot |G|,$$

which completes the proof.  $\square$

**12.K.14.** Solve the following system of polynomial equations in  $\mathbb{R}$ :

$$\begin{aligned} x^2 + yz + x, \\ z^2 + xy + z, \\ y^2 + xz + y. \end{aligned}$$

**Solution.** Using the algorithm 12.6.13 or appropriate software, we find the corresponding Gröbner basis for the lexicographic monomial ordering with  $x > y > z$ , consisting of six polynomials:

$$\begin{aligned} z^2 + 3z^3 + 2z^4, \\ z^2 + z^3 + 2yz^2 + 2yz^3, \\ y - yz - z - z^2 - 2yz^2 + y^2, \\ yz + z + z^2 + 2yz^2 + xz, \\ z^2 + xy + z, x^2 + yz + x. \end{aligned}$$

The roots of the first polynomial are  $z = 0, -1, -\frac{1}{2}$ . Discussing the individual cases, we find out that the system is satisfied exactly by the points  $(0, 0, 0), (-1, 0, 0), (0, -1, 0), (0, 0, -1)$  and  $(-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2})$ .  $\square$

It is recommended to think out thoroughly how useful these propositions are for solving combinatorial problems, c.f. 12.G.1 through 12.G.7.

Next, we shall come back to fields and after knowing the complete classification of finite Boolean algebras and finite abelian groups, we focus on finite fields.

As first step, we have to draw some elementary results on rings and univariate polynomial rings.

**12.4.16. Rings, ideals, quotients.** We are meeting rings on our journey permanently since the very beginning. Now we shall add some structural understanding of objects and homomorphisms in this category.

Similarly to groups, the inclusion of a subring  $K \subset L$  allows us to construct the residue class space  $L/K = \{a + K; a \in L\}$ , i.e., we consider the (commutative) additive group structure only. If we want to equip  $L/K$  by a ring structure applying the operations to representatives, we do not face problems with the addition (since all subgroups are normal in the abelian case), but we need the properties  $a \cdot K \subset K$  and  $K \cdot a \subset K$  for all  $a \in L$  to get the multiplication right, too. Indeed,  $(a+K) \cdot (b+K) = a \cdot b + a \cdot K + K \cdot b + K = a \cdot b + K$  under this assumption. Such a subring  $K \subset L$  is called an *ideal* in  $L$ , and we call  $L/K$  the *residue class ring*, or *quotient ring*. The verification of all ring properties for the addition and multiplication defined by means of the representatives is straightforward.

For all homomorphisms of rings,  $f : K \rightarrow L$ , the kernel  $\ker f = \{a \in K; f(a) = 0\}$  is clearly an ideal, since all multiples of zero are zero. Similarly to groups, the ideals  $K \subset L$  come with the short exact sequences of rings

$$(1) \quad 0 \rightarrow K \rightarrow L \rightarrow L/K \rightarrow 0.$$

By the very definition, an intersection of ideals is an ideal again. Thus we may talk about ideals generated by subsets in  $L$ . In commutative rings  $L$ , we distinguish the ideals generated by a single element and call them the *principal ideals*,  $\langle a \rangle = \{b \cdot a; b \in L\}$ . If there are no proper ideals (i.e. different from the trivial ones) except of the principal ones, we call  $L$  a *principal ideal domain*.

The integer ring  $\mathbb{Z}$  is a principal ideal domain. Indeed, if  $K \subset \mathbb{Z}$  and  $n$  is the smallest positive integer in  $K$ , then for any  $m \in K$ , the division yields  $m = qn + r$  with the remainder again in  $K$ . Thus  $r = 0$  (it cannot be smaller than  $n$ ) and we have proved  $K = \langle n \rangle$ . If  $n = 1$ , then  $K = \mathbb{Z}$ . The quotient ring  $\mathbb{Z}/\langle n \rangle$  is the ring of residue classes  $\mathbb{Z}_n$ .

Notice, there are no proper ideals in any field  $\mathbb{K}$ .

**12.4.17. Characteristic of rings and fields.** Consider a ring  $K$ . Its additive structure provides the cyclic subgroups generated by single elements  $a \in K$ ,  $G_a = \{a, a + a, 3a, \dots\}$ . If there is the minimal positive integer  $n$ , such that  $na = 0$ , then we say that the order of  $a$  is  $n$ . If there is such a (minimal)  $n$  valid for all  $a \in K$ , we say that the *characteristic* of the ring  $K$  is  $n$ . If there are elements of infinite order, or their

orders are not bounded, we say that the characteristic of  $K$  is infinity.

Clearly, the characteristic of  $\mathbb{Z}_n$  is  $n$ , while the scalars  $\mathbb{Z}$ ,  $\mathbb{Q}$ ,  $\mathbb{R}$ ,  $\mathbb{C}$  all have characteristic  $\infty$ .

**Lemma.** *If  $K$  is a finite ring with  $|K| = m$  elements, then the characteristic of  $K$  is a divisor of  $m$ .*

**PROOF.** The orders of the cyclic subgroups  $G_a$  must be finite (if  $ka = \ell a$ , then  $(k - \ell)a = 0$ , and there are only finitely many elements in  $K$ ) and they must be divisors of  $m$  (as known for all finite abelian groups).  $\square$

Another nice observation says, that finite integral domains are actually fields:

**Theorem.** *If  $K$  is a finite commutative ring without divisors of zero, then  $K$  is a field. Moreover the number of elements  $|K| = m$  is a power of a prime number,  $m = p^n$ , where  $p$  is the characteristic of  $K$ .*

**PROOF.** Let  $p$  be the characteristic of  $K$ . Suppose,  $p = k\ell$  is not a prime (clearly,  $1 < k \leq \ell < p$ ). Then looking at  $1 \in K$ ,  $0 = p \cdot 1 = (k\ell) \cdot 1 = (k \cdot 1) \cdot (\ell \cdot 1)$  and this is impossible without divisors of zero. Thus the characteristic of  $K$  is a prime  $p$ .

Next, for any  $0 \neq a \in K$ , all the elements  $a \cdot b$ ,  $b \in K$  must be different, since  $a \cdot b_1 = a \cdot b_2$  would imply  $a \cdot (b_1 - b_2) = 0$ , i.e.,  $b_1 = b_2$ . In particular, the unit 1 is expressed as  $1 = a \cdot b$  for some  $b$  and so  $b$  is the inverse to  $a$ . Thus  $K$  is a finite field.

Finally, the additive subgroup  $G_1 = \{1, 2 \cdot 1, \dots, p \cdot 1\}$  is closed under the multiplication, too (and is isomorphic to the field  $\mathbb{Z}_p$ ). Now,  $K$  can be viewed as a vector space over  $G_1$  and it must be generated by finite number of elements  $a_1, \dots, a_n$  (with big freedom in the choices, but always the same number, as known from the elementary linear algebra). But then it is clear that  $|K| = p^n$ , as requested.  $\square$

The theorem suggests that all finite fields are built from the simplest bricks — the so called *Galois fields*  $GF(p)$  isomorphic to  $\mathbb{Z}_p$  with prime  $p$ . They are often denoted  $\mathbb{F}_p$  in literature.

**12.4.18. Univariate polynomials over fields.** The ring of univariate polynomials  $K[x]$  over any field  $K$  is a principal ideal domain. Indeed, the same argument as for integers applies. Given a non-zero ideal  $J$ , there must be a polynomial  $f$  of lowest degree in  $J$ . If the degree is zero,  $J = K[x]$ . Otherwise, each  $g \in J$  is expressed as  $g = qf + r$  applying the division with remainder (cf. 12.3.6). Then  $r \in J$  and it must be of lower degree than  $f$  (which is impossible) or zero. Thus,  $J = \langle f \rangle$ .

Recall that each ideal  $J = \langle f \rangle \subset K[x]$ ,  $\deg f > 0$ , provides the quotient ring  $L = K[x]/J$ .

**Lemma.** *The ring  $L = K[x]/\langle f \rangle$ ,  $d = \deg f > 0$  contains  $K$  as subfield, and  $L$  is a vector space over  $\mathbb{K}$  of dimension at most  $d$ .*



PROOF. Since  $d = \deg f > 0$ , no constant polynomial can appear in  $\langle f \rangle$ . Thus  $a \mapsto a + J$  embeds  $K$  into  $L$  as a subring. Finally, the elements  $1 + J, x + J, \dots, x^{d-1} + J$  must generate  $L$  (remind we deal with univariate polynomials).  $\square$

**Proposition.** *The ring  $L = K[x]/\langle f \rangle$ ,  $\deg f > 0$  is a field if and only if the polynomial  $f$  does not have any roots in  $K$ . If  $K$  is finite, then  $|L| = p^{nm}$ , where  $p^n = |K|$  and  $m$  is the dimension of the vector space  $L$  over  $K$ .*

PROOF.  $\square$

In the situation of the proposition, we call  $L$  the *algebraic extension of the field  $K$*  with respect to the polynomial  $f$ . Clearly, the polynomial  $f$  becomes completely reducible,  $f = a(x - \alpha_1) \dots (x - \alpha_d)$ , over  $L$ . Thus the construction adds all the missing roots  $\alpha_i$  to  $\mathbb{K}$  and extends the field this way.

Considering a polynomial

**12.4.19. Theorem** (Classification of finite fields). *For each prime number  $p$  and power  $n \in \mathbb{N}$ , there is a field with  $p^n$  elements. Such field  $F$  is obtained as extension. All fields with the same number of elements are isomorphic.*

PROOF.  $\square$

**12.4.20. The fundamental theorem of algebra.** While it may happen that a polynomial over the real numbers has no roots, every polynomial over the complex numbers has a root. This is the statement of the so called *fundamental theorem of algebra*, which is presented here with an (essentially) complete proof. By this result, every polynomial in  $\mathbb{C}[x]$  has as many roots (including multiplicity) as its degree  $\deg f = k$ . Hence it always admits a factorization of the form

$$f(x) = b(x - a_1) \cdot (x - a_2) \dots (x - a_k)$$

for the complex roots  $a_i$  and the appropriate leading coefficient  $b$ .

**Theorem.** *The field  $\mathbb{C}$  is algebraically closed, i.e. every polynomial of degree at least one has a root.*

PROOF. We are going to provide an elementary proof based on simple real analysis, in particular the concept of continuity (the readers should be familiar with the techniques developed in Chapters 5 and 6).

Suppose that  $f \in \mathbb{C}[z]$  is a non-zero polynomial with no root, i.e.  $f(z) \neq 0$  for all  $z \in \mathbb{C}$ . Consider the mapping defined by

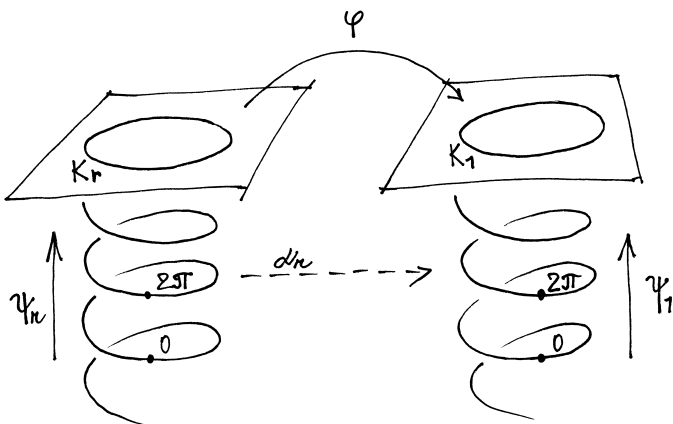
$$\varphi : \mathbb{C} \rightarrow \mathbb{C}, \quad z \mapsto \frac{f(z)}{|f(z)|}.$$

Then  $\varphi$  maps the entire  $\mathbb{C}$  into the unit circle

$$K_1 = \{e^{it}, t \in \mathbb{R}\} \subseteq \mathbb{C}.$$

By the assumption that  $f(z)$  is never zero, this mapping is well-defined. Next, we shall consider the restrictions of  $\varphi$  to the individual circles  $K_r \subseteq \mathbb{C}$  with center at zero and radius  $r \geq 0$





We can parameterize these circles by the mappings

$$\psi_r : \mathbb{R} \rightarrow K_r, \quad t \mapsto \psi_r(t) = r e^{it}.$$

For all  $r$ , the composition  $\kappa : (0, \infty) \times \mathbb{R} \rightarrow K_1$ ,  $\kappa(r, t) = \varphi \circ \psi_r(t)$ , is continuous in both  $t$  and  $r$ . Thus, for each  $r$ , there exists a mapping  $\alpha_r : \mathbb{R} \rightarrow \mathbb{R}$  which is uniquely given by the conditions  $0 \leq \alpha_r(0) < 2\pi$  and  $\kappa(r, t) = e^{i\alpha_r(t)}$ . Again, the obtained mapping  $\alpha_r$  continuously depends on  $r$ . Altogether, there is a continuous mapping

$$\alpha : \mathbb{R} \times (0, \infty) \rightarrow \mathbb{R}, \quad (t, r) \mapsto \alpha_r(t).$$

It follows from its construction that for every  $r$ ,  $\frac{1}{2\pi}(\alpha_r(2\pi) - \alpha_r(0)) = n_r \in \mathbb{Z}$ . Since  $\alpha$  is continuous in  $r$ , it means that  $n_r$  is an integer constant which is independent of  $r$ .

In order to complete the proof, it suffices to note that if  $f = a_0 + \dots + a_d z^d$  and  $a_d \neq 0$ , then for small values of  $r$ ,  $\alpha_r$  behaves nearly as a constant mapping, while for large values of  $r$ , it behaves almost as if  $f = z^d$ . First, calculate  $n_r$  for  $f = z^d$ , then make this statement more precise. This completes the proof.

The complex functions  $z \mapsto z^d$ ,  $z \mapsto \frac{z^d}{|z^d|}$  can be expressed easily using the trigonometric form of the complex numbers  $z = r(\cos \theta + i \sin \theta)$ :

$$\begin{aligned} z^d &= r^d (\cos d\theta + i \sin d\theta) = r^d e^{id\theta}, \\ \frac{z^d}{|z^d|} &= 1(\cos d\theta + i \sin d\theta) = e^{id\theta}. \end{aligned}$$

In this case, the mapping  $\varphi$  is simply a rotation of the complex plane, followed by the central projection onto the unit circle.

Then,  $\kappa_r(t) = e^{idt}$ , and so  $\alpha_r(t) = dt$ , regardless of  $r$ . It follows that  $n_r = d$  for the choice  $f = z^d$ . If  $f = az^d$  is chosen,  $a \neq 0$ , then there is no impact on the above result (verify this yourselves!).

Consider a general polynomial  $f = a_0 + \dots + a_d z^d$  with no root. Then  $a_0 \neq 0$  ( $a_0 = 0$ , implies that 0 is a root). For  $z \neq 0$ ,

$$\frac{f(z)}{a_d z^d} = 1 + \frac{1}{a_d} (a_0 z^{-d} + \dots + a_{d-1} z^{-1}).$$

Hence,  $\lim_{|z| \rightarrow \infty} \frac{f(z)}{a_d z^d} = 1$ . Knowing this, calculate

$$\begin{aligned} \lim_{|z| \rightarrow \infty} \left| \frac{f(z)}{|f(z)|} - \frac{a_d z^d}{|a_d z^d|} \right| &= \\ &= \lim_{|z| \rightarrow \infty} \left| \frac{f(z)}{a_d z^d} \frac{a_d z^d}{|a_d z^d|} \frac{|a_d z^d|}{|f(z)|} - \frac{a_d z^d}{|a_d z^d|} \right| = 0. \end{aligned}$$

Hence,  $n_r = d$  for large values of  $r$ .

A similar computation can be done for small values of  $r$ . Recall that  $a_0 \neq 0$ :

$$\frac{f(z)}{a_0} = 1 + \frac{1}{a_0}(a_1 z + \cdots + a_d z^d).$$

Thus,  $\lim_{|z| \rightarrow 0} \frac{f(z)}{a_0} = 1$ . In addition,  $\frac{f(z)}{|f(z)|} = \frac{f(z)}{a_0} \frac{a_0}{|a_0|} \frac{|a_0|}{|f(z)|}$ . Hence,  $\lim_{|z| \rightarrow 0} \frac{f(z)}{|f(z)|} = \lim_{|z| \rightarrow 0} \frac{a_0}{|a_0|}$ , i.e.  $n_r = 0$  for small values of  $r$ . Altogether, the degree of the polynomial is  $d = 0$ .  $\square$

## 5. Coding theory

It is often needed to transfer data while guaranteeing that they are transferred correctly. In some cases, it suffices to recognize whether the data is unchanged. In some cases, it can be transferred again. In other cases, this might not be reasonable, so that the data is needed to be recovered after a small number of errors in the transfer. This is the goal of coding theory. Some of the algorithms are explored now.

Notice that coding is quite different from encrypting. If no one but the addressee is meant to be able to read the message, then it should be encrypted. This topic is discussed briefly at the end of the previous chapter.

**12.5.1. Codes.** Data transfer is usually prone to errors. Since any information may be encoded as a sequence of *bits* (zeros or ones), the work is with  $\mathbb{Z}_2$  although the theory may be developed even for a general finite field. Furthermore, the length of the message to be transferred is assumed to be known in advance. Thus, one transfers  $k$ -bit words, where  $k \in \mathbb{N}$  is fixed.

It is desired to detect potential errors, and if possible, recover the original data. For this reason, further  $(n-k)$  bits are added to the  $k$ -bit word, where  $n$  is also fixed (and of course  $n > k$ ). These are called  $(n, k)$ -codes.

There are  $2^k$  binary words of length  $k$  and each should be mapped to one of the  $2^n$  possible *codewords*. For an  $(n, k)$ -code, there remain

$$2^n - 2^k = 2^k(2^{n-k} - 1)$$

words which are not codewords (if such a word is received, then an error has occurred). Thus, even for a large value of  $k$ , only a few bits added provide much redundant information.

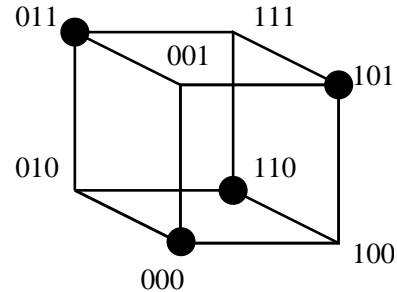
The simplest example is the *parity check code*. Having a message of length  $k$ , the codeword is created by adding a bit whose value is determined so that the total number ones would be even. This is an example of a  $(k+1, k)$ -code.

If there occur an odd number of errors during the transfer, then it can be detected with this simple code. Every two



codewords differ in at least two bits, but an error word differs from at least two codewords in only one bit. Therefore, this code is unable to recover the original message, even with the assumption that only one bit was changed.

The following diagram illustrates all 2-bit words with the parity bit added. The codewords are marked with a bold dot.



Moreover, the parity check code is unable to detect the error of interchanging a pair of adjacent bits, which often happens.

**12.5.2. Word distance.** In the diagram of the parity check (3, 2)-code, each error word is at the “same” distance from three codewords – those which differ in exactly bits. The other words are farther. Formally, this observation can be described by the following definition of distance:



WORD DISTANCE

The *Hamming distance* of a pair of words (of equal length) is the number of bits in which they differ.

Consider words  $x, y, z$  such that  $x$  and  $y$  differ in  $r$  bits, and  $y$  and  $z$  differ in  $s$  bits. Then,  $x$  and  $z$  differ in at most  $r + s$ , which verifies the triangle inequality for distances.

If the code is to detect errors in  $r$  bits, then the minimum distance between each pair of codewords must be at least  $r + 1$ . If the code is to recover errors in  $r$  bits, then there exists only one codeword whose distance from the received word is at most  $r$ . Thus, the following propositions are verified:

- Theorem.** (1) A code reliably detects at most  $r$  errors if and only if the minimum Hamming distance of the codewords is  $r + 1$ .
- (2) A code reliably detects and recovers at most  $r$  errors if and only if the minimum Hamming distance of the codewords is  $2r + 1$ .

**12.5.3. Construction of polynomial codes.** For practical applications, the codewords should be constructed efficiently so that they can be easily recognized among all the words. The parity check code is one example; another trivial possibility is to simply repeat the bits. For instance, the (3, 1)-code can be considered which triplicates each bit.



A systematic way for code construction is to use division of polynomials. A message  $b_0b_1 \dots b_{k-1}$  is understood as the polynomial

$$m(x) = b_0 + b_1x + \dots + b_{k-1}x^{k-1}$$

over the field  $\mathbb{Z}_2$ . The encoded message should be another polynomial  $v(x)$  of degree at most  $n - 1$ .

POLYNOMIAL CODES

Let  $p(x) = a_0 + \dots + a_{n-k}x^{n-k} \in \mathbb{Z}_2[x]$  be a polynomial with coefficients  $a_0 = 1, a_{n-k} = 1$ . The *polynomial code generated by a polynomial  $p(x)$*  is the  $(n, k)$ -code whose codewords are polynomial of degree less than  $n$  divisible by  $p(x)$ . A message  $m(x)$  is encoded as

$$v(x) = r(x) + x^{n-k}m(x),$$

where  $r(x)$  is the remainder in the division of the polynomial  $x^{n-k}m(x)$  by  $p(x)$ .

Check the claimed properties first. By the very definition of the codeword  $v(x)$  of a message  $m(x)$ :

$$\begin{aligned} v(x) &= r(x) + x^{n-k}m(x) = \\ &= r(x) + q(x)p(x) + r(x) = q(x)p(x), \end{aligned}$$

since the sum of two identical polynomials is always zero over  $\mathbb{Z}_2$ . Therefore, all codewords are divisible by  $p(x)$ .

On the other hand, if  $v(x)$  is divisible by  $p(x)$ , the above calculation can be read from right to left (setting  $r(x) = x^{n-k}m(x) - q(x)p(x)$ ), and so it is a codeword created by the above procedure.

From the definition, the codeword is created by adding  $n - k$  bits given by  $r(x)$  at the beginning of the word (and simply shifting the message to the right by that). It follows that the original message is contained in the polynomial  $v(x)$  and the decoding is easy.

Consider the two simple examples, already mentioned. First, note that  $p(x) = 1 + x$  divides  $v(x)$  if and only if  $v(1) = 0$ . This occurs if and only if  $v(x)$  has an even number of non-zero coefficients. So the polynomial  $p(x) = 1 + x$  generates the parity check  $(n, n - 1)$ -code for any  $n \geq 2$ .

Similarly, it is easily verified that the polynomial

$$p(x) = 1 + x + \dots + x^{n-1}$$

generates the  $(n, 1)$ -code of  $n$ -fold bit repetition. Dividing the polynomial  $b_0x^{n-1}$  by  $p(x)$ , gives the remainder  $b_0(1 + \dots + x^{n-2})$ , so the corresponding codeword is  $b_0p(x)$ .

**12.5.4. Error detection.** Let  $e(x)$  denote the *error vector*.

This is, the difference between the original message  $v \in (\mathbb{Z}_2)^n$  and the received data  $u$ :



$$u(x) = v(x) + e(x).$$

The error is detected if and only if the generator of the code (i.e. the polynomial  $p(x)$ ) does not divide  $e(x)$ . Therefore, polynomials over  $\mathbb{Z}_2[x]$  which do not happen to be divisors too often are of interest.

**Definition.** An irreducible polynomial  $p(x) \in \mathbb{Z}_2[x]$  of degree  $m$  is said to be *primitive* if and only if  $p(x)$  divides  $(1 + x^k)$  for  $k = 2^m - 1$ , but not for any smaller value of  $k$ .

**Theorem.** Let  $p(x)$  be a primitive polynomial of degree  $m$  and  $n \leq 2^m - 1$ . Then the polynomial  $(n, n - m)$ -code generated by  $p(x)$  detects all simple and double errors.

**PROOF.** If exactly one error occurs, then  $e(x) = x^i$  for some  $i, 0 \leq i < n$ . Since  $p(x)$  is irreducible, it cannot have a root in  $\mathbb{Z}_2$ . In particular, it cannot divide  $x^i$  since the factorization of  $x^i$  is unique. It follows that every simple error is detected.

If exactly two errors occur, then

$$e(x) = x^i + x^j = x^i (1 + x^{j-i})$$

for some  $0 \leq i < j < n$ .  $p(x)$  does not divide any  $x^i$ , and since it is primitive, it does not divide  $1 + x^{j-i}$  either, provided  $j-i < 2^m - 1$ . At the same time,  $p(x)$  is irreducible, which means that it does not divide the product  $e(x) = x^i(1 + x^{j-i})$ , which completes the proof.  $\square$

**12.5.5. Corollary.** Let  $q(x)$  be a primitive polynomial of degree  $m$  and  $n \leq 2^m - 1$ . Then the polynomial  $(n, n - m - 1)$ -code generated by the polynomial  $p(x) = q(x)(1 + x)$  detects all double errors as well as all errors with an odd number of bit flips.

**PROOF.** The codewords generated by the chosen polynomial  $p(x)$  are divisible by both  $x + 1$  and the primitive polynomial  $q(x)$ . As verified, the factor  $x + 1$  is responsible for parity checking, i.e. all codewords have an even number of non-zero bits. This detects any odd number of errors. By the above theorem, the second factor is able to detect all double errors.  $\square$

The following table illustrates the power of the above theorems for several primitive polynomials of low degrees. For instance, the last row says that by adding only 11 redundant bits to a message of length 1012, and employing the polynomial  $(x+1)p(x)$ , all single, double, triple, and odd-numbered errors in the transfer can be detected. These are already quite large numbers, with over 300 decimal digits.

primitive polynomial $p(x)$	redundant bits	codeword length
$1 + x$	1	1
$1 + x + x^2$	2	3
$1 + x + x^3$	3	7
$1 + x + x^4$	4	15
$1 + x^2 + x^5$	5	31
$1 + x + x^6$	6	63
$1 + x^3 + x^7$	7	127
$1 + x^2 + x^3 + x^4 + x^8$	8	255
$1 + x^4 + x^9$	9	511
$1 + x^3 + x^{10}$	10	1023

Note that quite strong results on the divisibility are used for the decomposition of polynomials derived in the second part of this chapter. But tools which would assist in constructing primitive polynomials are not mentioned.

Such tools come from the theory of finite fields. The name "primitive" reflects the connection to the primitive elements in the Galois fields  $G(2^m)$ . This theory also provides a convenient way of applying the Euclidean division, that is, of verifying whether or not the received word is a codeword, using the delayed registers. This is a simple circuit with as many elements as is the degree of the polynomial.<sup>12</sup>

**12.5.6. Linear codes.** Polynomial codes can also be described using elementary matrix calculus.



Recall that when working over the field  $\mathbb{Z}_2$ , caution is required when applying the results of elementary linear algebra, since then the property that  $v = -v$  implies  $v = 0$  is often used. This is not the case now.

However, the basic definition of vector spaces, existence of bases and descriptions of linear mappings by matrices are still valid. It is useful to recall the general theory and its applicability.

Start with a more general definition of codes, which only requires linear dependency of the codeword on the original message:

LINEAR CODES

Any injective linear mapping  $g : (\mathbb{Z}_2)^k \rightarrow (\mathbb{Z}_2)^n$  is a *linear code*. The  $k$ -by- $n$  matrix  $G$  that corresponds to this mapping (in the canonical bases) is called the *generating matrix* of the code.

For each message  $v$ , the corresponding codeword is given by

$$u = G \cdot v.$$

**Theorem.** *Every polynomial  $(n, k)$ -code is a linear code.*

**PROOF.** Use elementary properties of Euclidean division. Apply the assignment of the polynomial  $v(x) = r(x) + x^{n-k}m(x)$  determined by the original message  $m(x)$  to the sum of two messages  $m(x) = m_1(x) + m_2(x)$ . The remainder in the division  $x^{n-k}(m_1(x) + m_2(x))$  is, by uniqueness, given as the sum  $r_1(x) + r_2(x)$  of the remainders for the individual messages. It follows that

$$v(x) = r_1(x) + r_2(x) + x^{n-k}(m_1(x) + m_2(x)),$$

which is the desired additivity. Since the only non-zero scalar in  $\mathbb{Z}_2$  is 1, the linearity of the mapping of the message  $m(x)$  to the longer codeword  $v(x)$  is proved. Moreover, this mapping is clearly injective, since the original message  $m(x)$  is simply copied beyond the redundant bits.  $\square$

For instance, consider the  $(6, 3)$ -code generated by the polynomial  $p(x) = 1 + x + x^3$ , for encoding 3-bit words.

<sup>12</sup>More about the beautiful theory and its connection with codes can be found in the book Gilbert, W., Nicholson, K., Modern Algebra and its applications, John Wiley & Sons, 2nd edition, 2003, 330+xvii pp., ISBN 0-471-41451-4.

Evaluate it on the individual basis vectors  $m_i(x) = x^{i-1}$ ,  $i = 1, 2, 3$ , to get

$$\begin{aligned}v_0 &= (1 + x) + x^3, \\v_1 &= (x + x^2) + x^4, \\v_2 &= (1 + x + x^2) + x^5.\end{aligned}$$

It follows that the generating matrix of this  $(6, 3)$ -code is

$$G = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Polynomial codes always copy the original message beyond the redundant bits. So the generating matrix can be split into two blocks  $P$  and  $Q$  consisting respectively of  $n - k$  and  $k$  rows. Then  $Q$  equals the identity matrix  $\mathbb{I}_k$ .

**12.5.7. Theorem.** *Let  $g : (\mathbb{Z}_2)^k \rightarrow (\mathbb{Z}_2)^n$  be a linear code with generating matrix  $G$ , written in blocks as*

$$G = \begin{pmatrix} P \\ \mathbb{I}_k \end{pmatrix}.$$

*Then, the mapping  $h : (\mathbb{Z}_2)^n \rightarrow (\mathbb{Z}_2)^{n-k}$  with the matrix*

$$H = (\mathbb{I}_{n-k} \quad P)$$

*has the following properties:*

- (1)  $\text{Ker } h = \text{Im } g$ ,
- (2) *a received word  $u$  is a codeword if and only if  $H \cdot u = 0$ .*

**PROOF.** The composition  $h \circ g : (\mathbb{Z}_2)^k \rightarrow (\mathbb{Z}_2)^{n-k}$  is given by the product of matrices (computing over  $\mathbb{Z}_2$ ):

$$H \cdot G = (\mathbb{I}_{n-k} \quad P) \cdot \begin{pmatrix} P \\ \mathbb{I}_k \end{pmatrix} = P + P = 0.$$

Hence it is proved that  $\text{Im } g \subseteq \text{Ker } h$ . Since the first  $n - k$  columns of  $H$  are basis vectors of  $(\mathbb{Z}_2)^{n-k}$ , the image  $\text{Im } h$  has the maximum dimension  $n - k$ , which means that this image contains  $2^{n-k}$  vectors. Vector spaces over  $\mathbb{Z}_2$  are finite commutative groups, so the formula relating the orders of subgroups and quotient groups from subsection 12.4.10 can be used, thus obtaining

$$|\text{Ker } h| \cdot |\text{Im } h| = |(\mathbb{Z}_2)^n| = 2^n.$$

Therefore, the number of vectors in  $\text{Ker } h$  is equal to  $2^n \cdot 2^{k-n} = 2^k$ . In order to complete the proof of the first proposition, it suffices to note that the image  $\text{Im } g$  also has  $2^k$  elements.

The second proposition is a trivial corollary of the first one.  $\square$

The matrix  $H$  from the theorem is called the *parity check matrix* of the corresponding linear  $(n, k)$ -code.



For instance, the matrix  $H = (1\ 1\ 1)$  is the parity check matrix for the parity check  $(3, 2)$ -code, encoding 2-bit words. It is easily obtained from the matrix

$$G = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

that generates this code.

For the  $(6, 3)$ -code mentioned above, the parity check matrix is

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

**12.5.8. Error-correcting codes.** As seen, transferring a message  $u$  gives the result

$$v = u + e.$$

Over  $\mathbb{Z}_2$ , this is equivalent to  $e = u + v$ .

It follows that if the error  $e$  to be detected fixed, all the received words determined by the correct codewords  $u$  fill one of the cosets in the quotient space  $(\mathbb{Z}_2)^n/V$ , where  $V$  is the vector subspace  $V \subseteq (\mathbb{Z}_2)^n$  of the codewords.

The mapping  $h : (\mathbb{Z}_2)^n \rightarrow (\mathbb{Z}_2)^{n-k}$  corresponding to the parity check matrix has  $V$  as its kernel. Therefore, it induces the injective linear mapping  $\tilde{h} : (\mathbb{Z}_2)^n/V \rightarrow (\mathbb{Z}_2)^{n-k}$ . Clearly, the value  $\tilde{h}(v + V)$  on the coset generated by  $v$  is determined uniquely by the value  $H \cdot v$ .

#### SYNDROMES

The expression  $H \cdot v$ , where  $H$  is the parity check matrix for the considered linear code, is called the *syndrome* of  $v$ .

The following claim is a direct corollary of the construction and the above observations:

**Theorem.** *Two words are in the same class  $u + V$  if and only if they have the same syndromes.*

It follows that self-correcting codes can be constructed by choosing, for every syndrome, the element of the corresponding coset which is most likely to be the sent codeword. Naturally, when choosing the code, it is desirable to maximize the probability that it can correct single errors (and possibly even more errors).

Try it on the example of the  $(6, 3)$ -code for which the matrices  $G$  and  $H$  are already computed. Build the table of all syndromes and the corresponding words.

The syndrome 000 is possessed exactly by the codewords. All words with a given different syndrome are obtained by choosing one of them and adding all the proper codewords.

The following two tables display the syndromes in the first rows; the second rows then display the vector which has the least number of ones among the vectors of the corresponding coset. In almost all cases, this is just one value one there; only in the last column, there are two ones, and the element

is chosen where the ones are adjacent (because, for instance, multiple errors are more likely to be adjacent).

000	100	010	001
<b>000000</b>	<b>100000</b>	<b>010000</b>	<b>001000</b>
110100	010100	100100	111100
011010	111010	001010	010010
111001	011001	101001	110001
101110	001110	111110	100110
001101	101101	011101	000101
100011	000011	110011	101011
010111	110111	000111	011111

110	011	111	101
<b>000100</b>	<b>000010</b>	<b>000001</b>	<b>000110</b>
110000	110110	110101	110010
011110	011000	011011	011100
111101	111011	111000	111111
101010	101100	101111	101000
001001	001111	001100	001011
100111	100001	100010	100101
010011	010101	010110	010001

All the columns in the tables are affine subspaces whose modelling vector spaces are always the first column of the first table. This is because the code is linear, so that the set of all codewords forms a vector space, and the individual cosets of the quotient space are consequently affine subspaces.

In particular, the difference of each pair of words in the same column is a codeword. The words in bold are the leading representatives of the coset (affine space) that correspond to the given syndromes. These are the words with the least number of ones in their column. They correspond to the least number of bit flips which must be made to any word in the given column in order to get a codeword.

For instance, if a word 111101 is received, compute that its syndrome is 110. The leading representative of the coset of this syndrome is 000100. Subtract it from the received word, to obtain the codeword 111001. This is the codeword with the least Hamming distance from the received word. So the original message is most likely to be 001.

### 6. Systems of polynomial equations

In practical problems, objects or actions are often encountered which are described in terms of polynomials or systems of polynomial equations. For instance, the set of points in  $\mathbb{R}^3$  defined by two equations  $x^2 + y^2 - 1 = 0$  and  $z = 0$  is the circle which is centered at  $(0, 0, 0)$ , has radius 1 and lies in the plane  $xy$ .



Similarly, equations  $xz = 0$  and  $yz = 0$  considered in  $\mathbb{R}^3$  define the union of the line  $x = 0, y = 0$  and the plane  $z = 0$ . Notice we have to specify the space carefully, since  $x^2 + y^2 = 0$  defines a circle in  $\mathbb{R}^2$ , but it is a cylinder if viewed in  $\mathbb{R}^3$ .

Deciding whether or not a given point lies within a given body, finding extrema of algebraically defined subsets of multidimensional spaces, analyzing movements of parts of some machine, etc. are examples of such problems.

**12.6.1. Affine varieties.** For the sake of simplicity (existence of roots of polynomials), the work is mainly with the field of complex numbers. Some ideas are extended to the case of a general field  $\mathbb{K}$ .

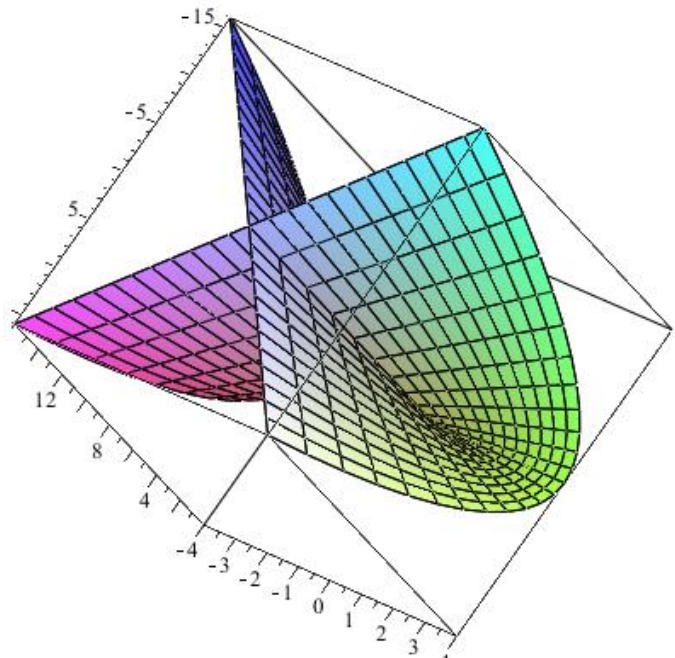
An affine  $n$ -dimensional space over a field  $\mathbb{K}$  is understood to be  $\mathbb{K}^n = \underbrace{\mathbb{K} \times \cdots \times \mathbb{K}}_n$  with the standard affine structure (see the beginning of Chapter 4).

As seen, a polynomial  $f = \sum_{\alpha} a_{\alpha} x^{\alpha} \in \mathbb{K}[x_1, \dots, x_n]$  can be viewed naturally as a mapping  $f: \mathbb{K}^n \rightarrow \mathbb{K}$ , defined by

$$f(u_1, \dots, u_n) := \sum_{\alpha} a_{\alpha} u^{\alpha}, \text{ where } u^{\alpha} = u_1^{\alpha_1} \cdots u_n^{\alpha_n}.$$

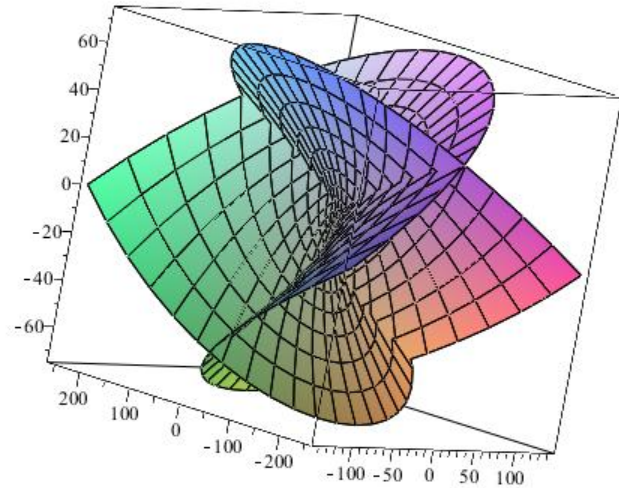
In dimension  $n = 1$ , the equality  $f(x) = 0$  describes only finitely many points of  $\mathbb{K}$ . In higher dimension, the similar equality  $f(x_1, \dots, x_n) = 0$  describes subsets similar to curves in the plane or surfaces in the space. However, they may be of quite complicated and self-intersecting shapes.

For instance, the set given by the equation  $(x^2 + y^2)^3 - 4x^2y^2 = 0$  look as a *quatrefoil* (see the illustration in the beginning of the part J in the other column). Another illustration of a two-dimensional surface is given by *Whitney's umbrella*  $x^2 - y^2z = 0$ , which, besides the part shown in the diagram, also includes the line  $\{x = 0, y = 0\}$ .



The diagram was drawn using the parametric description  $x = uv, y = v, z = u^2$ , whence the implicit description  $x^2 - y^2z = 0$  is easily guessed.

In the following illustration, there is the *Enneper surface* with parametrization  $x = 3u + 3uv^2 - u^3$ ,  $y = 3v + 3u^2v - v^3$ ,  $z = 3u^2 - 3v^2$ .



It is hard to imagine how to obtain the implicit description from this parametrization in hand. Nevertheless, there is an algorithm to eliminate the variables  $u$  and  $v$  from these three equations.

Quite a complex theory is needed to be developed for that. As usual, begin by formalization of the objects of interest.

#### AFFINE VARIETIES

Let  $f_1, \dots, f_s \in \mathbb{K}[x_1, \dots, x_n]$ . The *affine variety* in  $\mathbb{K}^n$  corresponding to the set of polynomials  $f_1, \dots, f_s$  is the set

$$\mathfrak{V}(f_1, \dots, f_s) = \{(a_1, \dots, a_n) \in \mathbb{K}^n; \\ f_i(a_1, \dots, a_n) = 0, i = 1, \dots, s\}$$

5

Affine varieties include conic sections, quadrics, and hyperquadrics, both singular and regular. Many curves and surfaces can be easily described as affine varieties.

The variety corresponding to a set of polynomials is the intersection of the varieties corresponding to the individual polynomials. For instance,  $\mathfrak{V}(x^2 + y^2 - 1, z) \subset \mathbb{R}^3$  is the circle which is centered at  $(0, 0, 0)$ , has radius 1 and lies in the plane  $xy$ .

Similarly,  $\mathfrak{V}(xz, yz) \subset \mathbb{R}^3$  is the union of the line  $x = 0$ ,  $y = 0$  and the plane  $z = 0$ , since it is exactly the points of these two objects where both the polynomials  $xz, yz$  are zero.

These examples illustrate that it is not easy to deal with the concept of dimension. Is the mentioned line, added to the plane, enough for the variety to be considered three-dimensional, or should one keep considering it two-dimensional with a certain anomaly?

Verify the following straightforward proposition:

**Theorem.** Let  $V = \mathfrak{V}(f_1, \dots, f_s)$  and  $W = \mathfrak{V}(g_1, \dots, g_t) \subseteq \mathbb{K}^n$  be affine varieties. Then,  $V \cup W$  and  $V \cap W$  are also affine varieties, where

$$V \cap W = \mathfrak{V}(f_1, \dots, f_s, g_1, \dots, g_t),$$

$$V \cup W = \mathfrak{V}(f_i g_j) \quad \text{for } 1 \leq i \leq s, 1 \leq j \leq t.$$

In the following subsections, some questions which arise in the context of varieties are answered:

Q1. Is the set  $\mathfrak{V}(f_1, \dots, f_s)$  empty?

Q2. Is the set  $\mathfrak{V}(f_1, \dots, f_s)$  finite?

Q3. How to understand the concept of dimension for varieties?

All of these problems can be “reasonably” solved for varieties over the complex numbers (as well as over any algebraically closed field); it is more difficult for the real numbers and nearly impossible for general fields. For instance, over the rational numbers, the question whether  $\mathfrak{V}(x^n + y^n - z^n) = \emptyset$  leads to the well-known Fermat’s last theorem, many times mentioned in Chapter 11.

**12.6.2. Parametrization.** For some purely practical operations with varieties, it is convenient to use the implicit representation (the one used so far). For instance, deciding whether a given point lies in a given variety, or inside the space enclosed by it, is quite easy using the implicit description. On the other hand, the parametric description may also come in handy in many situations (for example, it was used to draw the diagram).

The variety  $\mathfrak{V}(x + y + z - 1, x + 2y - z - 3)$  is a line (the intersection of two planes). If the system

$$\begin{aligned} x + y + z - 1 &= 0, \\ x + 2y - z - 3 &= 0, \end{aligned}$$

is solved, the parametric description of this line is immediate:

$$\begin{aligned} x &= -1 - 3t, \\ y &= 2 - 2t, \\ z &= t \end{aligned}$$

as known from the affine geometry. One needs to be more careful in general:

#### RATIONAL PARAMETRIZATION

**Definition.** A *rational parametric representation* of a variety  $\mathfrak{V}(f_1, \dots, f_r) \subseteq \mathbb{K}^n$  is a set of rational functions  $r_1, \dots, r_n \in \mathbb{K}(t_1, \dots, t_s)$  such that:

- the entire image of the mapping  $r = (r_1, \dots, r_n)$  is contained in  $\mathfrak{V}(f_1, \dots, f_r)$ ;
- $\mathfrak{V}(f_1, \dots, f_r)$  is the minimal affine variety which contains these points  $(x_1, \dots, x_n)$ .

Note that the parametrization is not required to describe all the points of the variety. This is important, as seen by a simple example of parametrization of a circle in the plane:

$$x = \frac{2t}{1+t^2}, \quad y = \frac{-1+t^2}{1+t^2},$$

which can be obtained using the stereographic projection. (Verify this in detail!) Note that this parametrization describes all points except for the point  $(0, 1)$ , from which we project, since this point is not reachable for any value of the parameter  $t$ . This is nobody's fault; it follows from the different topological properties of the circle and the line that there exists no global bijective rational parametrization.

In this connection, two more questions arise:

- Q4. *Does there exist a parametrization of a given variety and how to find it?*  
 Q5. *Is there an implicit description of a parametrically defined variety?*

The general answer to question 4 is negative. In fact, most affine varieties cannot be parametrized; or at least there is no algorithm for parametrization of the implicit description.

It is clear at first sight that a given variety may admit many implicit and parametric descriptions. In the case of implicit descriptions, it is given by representation in terms of several "generating" polynomials, and there is clearly much freedom in their choice. Once a parametrization is found, it can be composed with any rational bijection on the parameters in order to obtain another one.

**12.6.3. Ideals.** In order to avoid the dependence on the chosen equations that define a variety, consider all consequences of the given equations. This leads to the following algebraic concept of subsets in rings (which is similar to normal subgroups):



IDEALS

**Definition.** Let  $\mathbb{K}$  be a commutative ring. A subset  $I \subseteq \mathbb{K}$  is called an *ideal* if and only if  $0 \in I$  and

$$\begin{aligned} f, g \in I &\implies f + g \in I, \\ f \in I, h \in \mathbb{K} &\implies f \cdot h \in I. \end{aligned}$$

Since the definition contains only universal quantifiers (the properties are required for all elements in  $\mathbb{K}$  or  $I$ ), the intersection of two ideals is also an ideal. Consequently, ideals can be considered *generated* by subsets. Use the notation  $I = \langle a_1, \dots, a_n \rangle$ . It is easy to prove that such an ideal is

$$I = \left\{ \sum_i a_i b_i, b_i \in \mathbb{K} \right\},$$

where only finite sums are considered. (Check that this is the intersection of all ideals containing the set of the generators!) The set of generators may be infinite, too. If there are only finitely many generators, the ideal is said to be *finitely generated*.

It is easy to verify that each variety defines an ideal in the ring of polynomials in the following way:

## THE IDEAL OF A VARIETY

For a variety  $V = \mathfrak{V}(f_1, \dots, f_s)$ , set

$$\mathfrak{I}(V) := \{f \in \mathbb{K}[x_1, \dots, x_n]; \\ f(a_1, \dots, a_n) = 0, \forall (a_1, \dots, a_n) \in V\}.$$

**Lemma.** Let  $f_1, \dots, f_s, g_1, \dots, g_t \in \mathbb{K}[x_1, \dots, x_n]$  be polynomials. Then:

(1) if  $\langle f_1, \dots, f_s \rangle = \langle g_1, \dots, g_t \rangle$ , then

$$V = \mathfrak{V}(f_1, \dots, f_s) = \mathfrak{V}(g_1, \dots, g_t);$$

(2)  $\mathfrak{I}(V)$  is an ideal, and  $\langle f_1, \dots, f_s \rangle \subseteq \mathfrak{I}(V)$ .

**PROOF.** If a point  $a = (a_1, \dots, a_n)$  lies in a variety  $\mathfrak{V}(f_1, \dots, f_s)$ , then any polynomial of the form

$$f = h_1 f_1 + \dots + h_s f_s$$

(i.e. any member of the ideal  $I = \langle f_1, \dots, f_s \rangle$ ) takes zero at  $a$ . In particular, this means that all the polynomials  $g_i$  take zero at  $a$ . Hence

$$\mathfrak{V}(f_1, \dots, f_s) \subseteq \mathfrak{V}(g_1, \dots, g_t).$$

The other inclusion can be proved similarly.

In order to verify the second proposition, choose  $g, g' \in \mathfrak{I}(V)$ ,  $h \in \mathbb{K}[x_1, \dots, x_n]$ . Then, for any point  $a \in V$  that

$$(gh)(a) = 0 \Rightarrow gh \in \mathfrak{I}(V),$$

$$(g + g')(a) = 0 \Rightarrow g + g' \in \mathfrak{I}(V).$$

Hence  $\mathfrak{I}(V)$  is an ideal in  $\mathbb{K}[x_1, \dots, x_n]$ .

For any polynomial  $f = h_1 f_1 + \dots + h_s f_s \in \langle f_1, \dots, f_s \rangle$  and a point  $a \in V$ ,  $f(a) = 0$ , which proves the desired inclusion.  $\square$

The simplest examples are trivial varieties – a single point and the entire affine space:

$$\mathfrak{I}(\{(0, 0, \dots, 0)\}) = \langle x_1, \dots, x_n \rangle,$$

$$\mathfrak{I}(\mathbb{K}^n) = \{0\} \quad \text{for any infinite field } \mathbb{K}.$$

The other inclusion of the second part of the lemma does not hold in general. For instance, the variety  $\mathfrak{V}(x^2, y^2)$  contains only the single point –  $(0, 0)$ . This means that  $\mathfrak{I}(V) = \langle x, y \rangle \supset \langle x^2, y^2 \rangle$ .

If  $V, W \subseteq \mathbb{K}^n$  are varieties, then

$$V \subseteq W \implies \mathfrak{I}(V) \supseteq \mathfrak{I}(W).$$

In other words, a polynomial which takes zero at each point of a given variety clearly takes zero at each point of any of the variety's subsets.

Now, further natural problems can be formulated:

Q6. Is every ideal  $I \in \mathbb{K}[x_1, \dots, x_n]$  finitely generated?

Q7. Is there an algorithm which decides whether  $f \in \langle f_1, \dots, f_s \rangle$ ?

Q8. What is the precise relation between  $\langle f_1, \dots, f_s \rangle$  and  $\mathfrak{I}(\mathfrak{V}(f_1, \dots, f_s))$ ?

**12.6.4. Dimension 1.** Consider univariate polynomials first

$$f = a_0x^n + a_1x^{n-1} + \cdots + a_n, \quad \text{where } a_0 \neq 0.$$

The leading term of a polynomial is defined to be  $LT(f) := a_0x^n$ . Clearly,

$$\deg f \leq \deg g \iff LT(f)|LT(g).$$

Let  $\mathbb{K}$  be a field and  $g$  a non-zero polynomial. Every polynomial  $f \in \mathbb{K}[x]$  can be written in a unique way as

$$f = q \cdot g + r, \quad \text{where } r = 0 \text{ or } \deg r < \deg g.$$

In fact, the quotient  $q$  and the remainder  $r$  can be computed by the following algorithm:

- (1)  $q := 0, r := f$
- (2) while  $r \neq 0 \wedge LT(g)|LT(r)$ 
  - (a)  $q := q + LT(r)/LT(g)$
  - (b)  $r := r - LT(r)/LT(g) \cdot g$

When checking the loop condition, the invariant  $f = q \cdot g + r$  holds, so the algorithm answers correctly as soon as the loop condition becomes "false". Since the degree of  $r$  is decreasing, the algorithm eventually terminates.

**Corollary.** *Let  $\mathbb{K}$  be a field. Then, every ideal in the polynomial ring  $\mathbb{K}[x]$  is of the form  $\langle f \rangle$ .*

**PROOF.** Consider an ideal  $I \subseteq \mathbb{K}[x]$ . If  $I = \{0\}$ , then the ideal is generated by the zero polynomial. If  $I$  contains a non-zero polynomial  $f$ , then choose any with the lowest degree. Clearly  $\langle f \rangle \subseteq I$ .

For any polynomial  $g \in I$ , consider the Euclidean division of  $g$  by  $f$ , i.e.  $g = qf + r$ . Clearly,  $qf \in I$ , which means that  $r \in I$  as well. However, the degree of  $f$  is as small as possible, so  $r = 0$ . Therefore,  $g$  is a multiple of  $f$ , and  $I = \langle f \rangle$ .  $\square$

An ideal which is generated by a single element is called a *principal ideal*. A ring which has the property of the last lemma is called a *principal ideal domain*.

Recall the Euclidean algorithm for the greatest common divisor  $h = GCD(f, g)$  of polynomials  $f$  and  $g$  (the variable  $h$  contains the desired greatest common divisor when the algorithm terminates):

- (1)  $h := f, s := g$
- (2) while  $s \neq 0$ 
  - (a)  $r := h \bmod s$
  - (b)  $h := s$
  - (c)  $s := r$

Let  $f = q \cdot g + r$  and  $h = GCD(f, g)$ . Then,  $h|r, g$  and

$$\forall p \in \mathbb{K}[x]: p|r, g, \quad \text{so } p|f \text{ and } p|h.$$

Hence,  $h$  is  $GCD(r, g)$ . Trivially,  $GCD(h, 0) = h$ , so the algorithm computes  $GCD(f, g)$  correctly. Since the degree of  $r$  is strictly decreasing, the algorithm eventually terminates.

It follows that each pair of polynomials has a greatest common divisor. It is unique up to multiplication by a scalar. Indeed, if two polynomials are greatest common divisors of a given pair of polynomials, then they must divide each other,



which is, for polynomials, possible only in the case mentioned.

The greatest common divisor of more polynomials is defined as follows: If  $s > 2$ , then

$$GCD(f_1, \dots, f_s) := GCD(f_1, GCD(f_2, \dots, f_s)).$$

**Lemma.** *Let  $f_1, \dots, f_s$  be polynomials. Then,  $\langle GCD(f_1, \dots, f_s) \rangle = \langle f_1, \dots, f_s \rangle$ .*

**PROOF.**  $GCD(f_1, \dots, f_s)$  divides all the polynomials  $f_i$ . Hence the principal ideal  $\langle GCD(f_1, \dots, f_s) \rangle$  contains the ideal  $\langle f_1, \dots, f_s \rangle$ . The other inclusion follows immediately from Bezout's identity.  $\square$

Earlier, eight questions were formulated. Here are some answers for dimension 1:

- Since  $\mathfrak{V}(f_1, \dots, f_s) = \mathfrak{V}(GCD(f_1, \dots, f_s))$ , the problem of emptiness of a given variety reduces to the problem of existence of a root of a single polynomial.
- For the same reason, each variety is a finite set of isolated points – the roots of the polynomial  $GCD(f_1, \dots, f_s)$ , except for the case when  $GCD(f_1, \dots, f_s) = 0$ . This can happen only if  $f_1 = f_2 = \dots = f_s = 0$ , and then the variety is the entire  $\mathbb{K}$ .
- The concept of dimension is not of much interest in this case; each variety has dimension zero, being a discrete set of points.
- Each ideal can be generated by a single polynomial.
- $f \in \langle f_1, \dots, f_s \rangle \iff GCD(f_1, \dots, f_s) | f$ .
- Denoting  $\langle f \rangle := \mathfrak{I}(\mathfrak{V}(f_1, \dots, f_s))$ , then  $f$  and  $GCD(f_1, \dots, f_s)$  may differ only in root multiplicities.

**12.6.5. Monomial ordering.** In order to generalize the Euclidean division of polynomials for more variables, one must first find an appropriate analogy for the degree of a polynomial and its leading term.



The Euclidean division of a polynomial  $f \in \mathbb{K}[x_1, \dots, x_n]$  by polynomials  $g_1, \dots, g_s$  is to be an expression of the form

$$f = a_1 g_1 + \dots + a_s g_s + r$$

where no term of the remainder  $r$  is divisible by the leading term of any of the polynomials  $g_i$ .

Try this with  $f = x^2 y + x y^2 + y^2$ ,  $g_1 = xy - 1$ , and  $g_2 = y^2 - 1$ . The first division yields

$$f = (x + y) \cdot g_1 + (x + y^2 + y).$$

$LT(y^2 - 1)$  does not divide  $x$  (the leading term of the remainder), so, theoretically, continuation is not possible.

However,  $x$  can be moved into the remainder, thus obtaining the result

$$f = (x + y) \cdot g_1 + g_2 + (x + y + 1).$$

No term of the remainder is divisible by either  $LT(g_1)$  or  $LT(g_2)$ . How are the leading terms determined?

MONOMIAL ORDERING

A *monomial ordering* on  $\mathbb{K}[x_1, \dots, x_n]$  is a well-ordering (every non-empty subset has a least element)  $<$  on  $\mathbb{N}^n$  which satisfies

$$\forall \alpha, \beta, \gamma \in \mathbb{Z}^n : \alpha < \beta \implies \alpha + \gamma < \beta + \gamma.$$

An ordering on  $\mathbb{N}^n$  induces an ordering on monomials as soon as the order of the variables  $x_1 < x_2 < \dots < x_n$  is fixed.

Each polynomial can be rearranged as a decreasing sequence of monomials (ignoring the coefficients for now).

The following three definitions introduce the most common monomial orderings. Each ordering assumes that the order of the variables is fixed, usually  $x_1 > x_2 > \dots > x_n$ .

**Definition.** Let  $\alpha, \beta \in \mathbb{N}^n$ . The *lexicographic ordering*  $<_{\text{lex}}$  is defined by

$$\alpha >_{\text{lex}} \beta \iff \text{the left-most non-zero term of } \alpha - \beta \text{ is positive.}$$

The *graded lexicographic ordering*  $<_{\text{glex}}$  is defined by

$$\alpha >_{\text{glex}} \beta \iff |\alpha| > |\beta| \quad \text{or} \\ |\alpha| = |\beta| \quad \text{and } \alpha >_{\text{lex}} \beta.$$

The *graded reverse lexicographic ordering*  $<_{\text{grevlex}}$  is defined by

$$\alpha >_{\text{grevlex}} \beta \iff |\alpha| > |\beta| \quad \text{or} \\ |\alpha| = |\beta| \quad \text{and the right-most nonzero} \\ \text{term of } \alpha - \beta \text{ is negative.}$$

If  $x > y > z$ , then  $x >_{\text{grevlex}} y >_{\text{grevlex}} z$ , but  $x^2yz^2 >_{\text{glex}} xy^3z$ , yet  $x^2yz^2 <_{\text{grevlex}} xy^3z$ .

Verify in detail that  $>_{\text{lex}}, >_{\text{glex}}, >_{\text{grevlex}}$  are monomial orderings!

**12.6.6. Multivariate division with remainder.** Consider a non-zero polynomial  $f = \sum_{\alpha \in \mathbb{N}^n} a_{\alpha} x^{\alpha}$  in  $\mathbb{K}[x_1, \dots, x_n]$  and  $<$  a monomial ordering. Then define the *degree*, *leading coefficient*, *leading monomial*, and *leading term* of  $f$  as follows:

- $\text{multideg } f := \max\{\alpha \in \mathbb{N}^n, a_{\alpha} \neq 0\}$ ,
- $LC f := a_{\text{multideg } f}$ ,
- $LM f := x^{\text{multideg } f}$ ,
- $LT f := LC f \cdot LM f$ .

Of course, these concepts depend on the underlying monomial ordering.

**Lemma.** Let  $f, g \in \mathbb{K}[x_1, \dots, x_n]$  and  $<$  be a monomial ordering. Then,

- (1)  $\text{multideg}(f \cdot g) = \text{multideg } f + \text{multideg } g$ ,
- (2)  $f + g \neq 0 \implies \text{multideg}(f + g) \leq \max\{\text{multideg } f, \text{multideg } g\}$ .

PROOF. Both claims are straightforward corollaries of the definitions.  $\square$

**Theorem.** Let  $<$  be a monomial ordering and  $F = (f_1, \dots, f_s)$  be an  $s$ -tuple of polynomials in  $\mathbb{K}[x_1, \dots, x_n]$ . Then, every polynomial  $f \in \mathbb{K}[x_1, \dots, x_n]$  can be expressed as

$$f = a_1 f_1 + \dots + a_s f_s + r,$$

where  $a_i, r \in \mathbb{K}[x_1, \dots, x_n]$  for all  $i = 1, 2, \dots, s$ . Moreover, either  $r = 0$  or  $r$  is a linear combination of monomials none of which is divisible by any of  $LT f_1, \dots, LT f_s$ , and if  $a_i f_i \neq 0$ , then  $\text{multideg } f \geq \text{multideg } a_i f_i$  for each  $i$ .

The polynomial  $r$  is called the remainder of the multivariate division  $f/F$ .

**PROOF.** The theorem says nothing about uniqueness of the result. The following algorithm produces a possible solution and thus proves the theorem.

In the sequel, consider the output of this algorithm to be the result of the division.

- (1)  $a_1 := 0, \dots, a_s := 0, r := 0, p := f$
- (2) while  $p \neq 0$ 
  - (a)  $i := 1$
  - (b)  $d := \text{false}$
  - (c) while  $i \leq s \wedge \text{not } d$ 
    - (i) if  $LT f_i | LT p$ 

$$a_i := a_i + LT p / LT f_i$$

$$p := p - (LT p / LT f_i) \cdot f_i$$

$$d := \text{true}$$
    - (ii) else  $i := i + 1$
  - (d) if not  $d$ 
    - (i)  $r := r + LT p$
    - (ii)  $p := p - LT p$

In every iteration of the outer loop, exactly one of the commands 2(c)i, 2(d)ii is executed, so the degree of  $p$  decreases. Therefore, the algorithm eventually terminates.

When checking the loop condition, the invariant  $f = a_1 f_1 + \dots + p + r$  holds, and each term of each  $a_i$  is the quotient  $LT p / LT f_i$  from some moment. The degrees of these terms are less than the current degree of  $p$ , which is at most the degree of  $f$ . Altogether, the degree of each  $a_i f_i$  is at most the degree of  $f$ .  $\square$

In the ring  $\mathbb{K}[x_1, \dots, x_n]$ , the following implication clearly holds:

$$f = a_1 f_1 + \dots + a_s f_s + 0 \implies f \in \langle f_1, \dots, f_s \rangle.$$

However, the converse is generally not true for multivariate division:

Consider  $f = xy^2 - x$ ,  $f_1 = xy + 1$ ,  $f_2 = y^2 - 1$ . The algorithm outputs

$$f = y(xy + 1) + 0(y^2 - 1) + (-x - y),$$

but  $f = x(y^2 - 1)$ , so that  $f \in \langle f_1, f_2 \rangle$ .

The next goal is to find some distinguished generators of the ideals  $I = \langle f_1, \dots, f_s \rangle$  which would behave better. In a certain sense, this is a similar procedure to the Gaussian elimination of variables for systems of linear equations. Begin with some special assumptions about the ideals.

**12.6.7. Monomial ideals.** An ideal  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  is called *monomial* if and only if there is a set of multi-indices  $A \subseteq \mathbb{N}^n$  such that  $I$  is generated by the monomials  $x^\alpha$  with  $\alpha \in A$ .



This means that all polynomials in  $I$  are of the form  $\sum_{\alpha \in A} h_\alpha x^\alpha$ , where  $h_\alpha \in \mathbb{K}[x_1, \dots, x_n]$ .

Clearly, for a monomial ideal  $I$   $x^\beta \in I$  if and only if there exists an  $\alpha \in A$  such that  $x^\alpha$  divides  $x^\beta$ .

**Lemma.** Let  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  be a monomial ideal and  $f \in \mathbb{K}[x_1, \dots, x_n]$  a polynomial. Then, the following propositions are equivalent:

- (1)  $f \in I$ ,
- (2) each term of  $f$  lies in  $I$ ;
- (3) the polynomial  $f$  is a linear combination of monomials from  $I$  with coefficients from  $\mathbb{K}$ .

**PROOF.** The implications (3)  $\implies$  (2)  $\implies$  (1) are obvious. It remains to prove (1)  $\implies$  (3).

Write the polynomial  $f$  as  $f = \sum_{\alpha} a_\alpha x^\alpha$ , where  $a_\alpha \in \mathbb{K}$ . It follows from the assumption  $f \in I$  that  $f = \sum_{\beta \in A} h_\beta x^\beta$ , where  $x^\beta \in I$  and  $h_\beta \in \mathbb{K}[x_1, \dots, x_n]$ . Each term  $a_\alpha x^\alpha$  must equal some term of the other equality. Hence each term  $a_\alpha x^\alpha$  of the polynomial  $f$  can be expressed as the sum of expressions  $d x^{\beta+\delta}$ , where  $d \in \mathbb{K}$ ,  $x^\beta \in I$ . However, this means that  $x^\alpha \in I$ , so that (3) holds.  $\square$

**Corollary.** Two monomial ideals coincide if and only if they contain the same monomials.

The following theorem goes much further. It says that every monomial ideal is finitely generated and, moreover, the finite set of generators may be chosen from any given set of generators.

**12.6.8. Theorem** (Dickson's lemma). Every monomial ideal  $I = \langle x^\alpha, \alpha \in A \rangle \subseteq \mathbb{K}[x_1, \dots, x_n]$  can be written in the form  $I = \langle x^{\alpha_1}, \dots, x^{\alpha_s} \rangle$ , where  $\alpha_1, \dots, \alpha_s \in A$ .

**PROOF.** Proceed by induction on the number of variables. If  $n = 1$ , then  $I \subseteq \mathbb{K}[x]$ ,  $I = \langle x^\alpha, \alpha \in A \subseteq \mathbb{N} \rangle$ . The set of exponents in  $A$  has a minimum, so denote it by  $\beta := \min A$ . Then,  $x^\beta$  divides each monomial  $x^\alpha$  with  $\alpha \in A$ , so  $I = \langle x^\beta \rangle$ .



Now suppose  $n > 1$  and assume that the proposition is true for fewer variables. Denote the variables  $x_1, \dots, x_{n-1}, y$ , and write monomials in the form  $x^\alpha y^m$  where  $\alpha \in \mathbb{N}^{n-1}$ ,  $m \in \mathbb{N}$ . Suppose that  $I \subseteq \mathbb{K}[x_1, \dots, x_{n-1}, y]$  is monomial, and define  $J \subseteq \mathbb{K}[x_1, \dots, x_{n-1}]$  by

$$J := \langle x^\alpha, \exists m \in \mathbb{N}, x^\alpha y^m \in I \rangle.$$

Clearly,  $J$  is a monomial ideal in  $n - 1$  variables, so by the induction hypothesis,  $J = \langle x^{\alpha_1}, \dots, x^{\alpha_s} \rangle$ . It follows from the definition of  $J$  that there are minimal integers  $m_i \in \mathbb{N}$  such that  $x^{\alpha_i} y^{m_i} \in I_A$ . Denote  $m := \max\{m_i\}$  and define an analogous system of ideals  $J_k \subseteq \mathbb{K}[x_1, \dots, x_{n-1}]$  for  $0 \leq k \leq m - 1$

$$J_k := \langle x^\beta; x^\beta y^k \in I_A \rangle.$$

Again, all the ideals  $J_k$  satisfy the induction hypothesis, so they can be expressed as

$$J_k = \langle x^{\alpha_{k,1}}, \dots, x^{\alpha_{k,s_k}} \rangle.$$

It remains to show that  $I$  is generated by the following finite set of monomials:

$$\begin{aligned} & x^{\alpha_1} y^m, \dots, x^{\alpha_s} y^m, \\ & x^{\alpha_{0,1}} y^0, \dots, x^{\alpha_{0,s_0}} y^0, \\ & \vdots \\ & x^{\alpha_{m-1,1}} y^{m-1}, \dots, x^{\alpha_{m-1,s_{m-1}}} y^{m-1}. \end{aligned}$$

Consider a monomial  $x^\alpha y^p \in I$ . Either of the following cases occurs:

- $p \geq m$ . Then,  $x^\alpha \in J$ ,  $k = p$ , so one of  $x^{\alpha_1} y^m, \dots, x^{\alpha_s} y^m$  divides  $x^\alpha y^p$ .
- $p < m$ . Then, analogously,  $x^\alpha \in J_k$ , and one of  $x^{\alpha_{k,1}} y^k, \dots, x^{\alpha_{k,s_k}} y^k$  divides  $x^\alpha y^p$ .

By the previous lemma, each polynomial  $f \in I$  can be expressed as a linear combination of monomials from  $I$ . Each of these is divisible by one of the generators; hence  $f$  lies in the ideal generated by them. Therefore,  $I$  is a subset of that. The other inclusion is trivial, which completes the proof of Dickson's lemma.  $\square$

**12.6.9. Hilbert's theorem.** Everything is now at hand for the discussion of ideal bases in polynomial rings. The main idea is the maximal utilization of the information about the leading terms among the generating polynomials and in the ideal. For a non-zero ideal  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$ , denote



$$LT I := \{ax^\alpha; \exists f \in I: LT f = ax^\alpha\}.$$

Clearly,  $\langle LT I \rangle$  is a monomial ideal, so by Dickson's lemma,  $\langle LT I \rangle = \langle LT g_1, \dots, LT g_s \rangle$  for appropriate  $g_1, \dots, g_s \in I$ .

**Theorem.** Every ideal  $I \in \mathbb{K}[x_1, \dots, x_n]$  is finitely generated.

**PROOF.** The statement is trivial for  $I = \{0\}$ . So suppose  $I \neq \{0\}$ . By Dickson's lemma and the above note, there are  $g_1, \dots, g_s \in I$  such that  $\langle LT I \rangle = \langle LT g_1, \dots, LT g_s \rangle$ .

Clearly,  $\langle g_1, \dots, g_s \rangle \subseteq I$ . Choose any polynomial  $f \in I$  and divide it by the  $s$ -tuple  $g_1, \dots, g_s$ .

$$f = a_1 g_1 + \dots + a_s g_s + r,$$

is obtained where no term of  $r$  is divisible by any of  $LT g_1, \dots, LT g_s$ .

Since  $r = f - a_1 g_1 - \dots - a_s g_s$ ,  $r \in I$ , and also  $LT r \in LT I$ . This means that  $LT r \in \langle LT I \rangle$ . Admit that  $r \neq 0$ . Since  $\langle LT I \rangle$  is monomial,  $LT r$  must be divisible by one of its generators, i.e.  $LT g_1, \dots, LT g_s$ . This contradicts the result of the multivariate division algorithm. Therefore,  $r = 0$  and  $I$  is generated by  $g_1, \dots, g_s$ .  $\square$

**12.6.10. Gröbner bases.** The basis used in the proof of Hilbert's theorem has the properties stated in the following definition:

GRÖBNER BASES OF IDEALS

**Definition.** A finite set of generators  $g_1, \dots, g_s$  of an ideal  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  is called *Gröbner basis* if and only if  $\langle LT I \rangle = \langle LT g_1, \dots, LT g_s \rangle$ .

**Corollary.** Every ideal  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  has a Gröbner basis. Every set of polynomials  $g_1, \dots, g_s \in I$  such that  $\langle LT I \rangle = \langle LT g_1, \dots, LT g_s \rangle$  is a Gröbner basis of  $I$ .

**Example.** Return to the remark on similarity with the Gaussian variable elimination for systems of linear equations. That is, illustrate the general results above on the simplest case of polynomials of degree one with the lexicographic ordering.



Denote the generators  $f_i = \sum_j a_{ij}x_j + a_{i0}$ . Consider the matrix  $A = (a_{ij})$ , where  $i = 1, \dots, s$  and  $j = 0, \dots, n$ , and apply the Gaussian elimination to it. This gives a matrix  $B = (b_{ij})$  in echelon form. Zero rows can be omitted from it. Hence there is a new basis  $g_1, \dots, g_t$ , where  $t \leq s$ .

Due to the performed steps, each  $f_i$  can be expressed as a linear combination  $g_1, \dots, g_t$ , which means that

$$\langle f_1, \dots, f_s \rangle = \langle g_1, \dots, g_t \rangle.$$

Now, verify that these polynomials  $g_1, \dots, g_t$  form a Gröbner basis. Without loss of generality, assume that the variables are labeled so that  $LM g_i = x_i$  for  $i = 1, \dots, t$ . Any polynomial  $f \in I$  can be written as

$$f = h_1 f_1 + \dots + h_s f_s = h'_1 g_1 + \dots + h'_t g_t.$$

It is required that  $LT f \in \langle LT g_1, \dots, LT g_t \rangle$ . That is,  $LT f$  is divisible by one of  $x_1, \dots, x_t$ . Suppose that  $f$  contains only the variables  $x_{t+1}, \dots, x_n$ . However, then  $h'_1 = 0$ , since  $x_1$  is only in  $g_1$  by the echelon form of  $B$ . Analogously,  $h'_2 = \dots = h'_t = 0$ , and so  $f = 0$ .

The existence of the very special bases is now proved.



However, they cannot yet be constructed algorithmically. This is the goal of the following subsections.

**12.6.11. Theorem.** Let  $G = \{g_1, \dots, g_t\}$  be a Gröbner basis of an ideal  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  and  $f$  a polynomial in  $\mathbb{K}[x_1, \dots, x_n]$ . Then, there is a unique  $r = \sum_{\alpha} a_{\alpha} x^{\alpha} \in \mathbb{K}[x_1, \dots, x_n]$  such that:

- (1) no term of  $r$  is divisible by any of  $LT g_1, \dots, LT g_t$ , i.e.  $\forall \alpha \forall i: LT g_i \nmid a_{\alpha} x^{\alpha}$ ;
- (2)  $\exists g \in I: f = g + r$ .

**PROOF.** The algorithm for multivariate division produces

$$f = a_1 g_1 + \dots + a_t g_t + r,$$

where  $r$  satisfies the condition (1). Select  $g$  as  $a_1 g_1 + \dots + a_t g_t$ , which of course lies in  $I$ .

It remains to prove uniqueness. Suppose that

$$f = g + r = g' + r',$$

where  $r \neq r'$ . Clearly,  $r - r' = g' - g \in I$ . Since  $G$  is a Gröbner basis,  $LT(r - r')$  is divisible by one of  $LT g_1, \dots, LT g_t$ . There are two possibilities:

- $LM r \neq LM r'$ . The one with the higher degree must be divisible by one of  $LT g_1, \dots, LT g_t$ , which contradicts condition (1).
- $LM r = LM r'$  and  $LC r \neq LC r'$ . Then both the monomials  $LM r$  and  $LM r'$  must be divisible by one of  $LT g_1, \dots, LT g_t$ , which is again a contradiction.

It follows that  $LT r = LT r'$  and the inductive argument shows that  $r = r'$ .  $\square$

The previous theorem generalizes the Euclidean division, with an ideal instead of a divisor. In the univariate case, this is no generalization, since every ideal is generated by a single polynomial. If it is only the remainder which is of interest, the order of polynomials in the Gröbner basis does not matter. Hence it makes sense to define the notation  $\bar{f}^G$  for the remainder in the division  $f/G$ , provided  $G = (g_1, \dots, g_s)$  is a Gröbner basis.

**Corollary.** *Let  $G = \{g_1, \dots, g_t\}$  be a Gröbner basis of an ideal  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  and  $f$  a polynomial in  $\mathbb{K}[x_1, \dots, x_n]$ . Then,  $f \in I$  if and only if the remainder  $\bar{f}^G$  is zero.*

**12.6.12. Syzygies.** The next step is to find a sufficient “testing set” of polynomials of a given ideal which allows us to verify whether the considered system is a Gröbner basis. Again, we wish to test this by means of multivariate division only.



For  $\alpha = \text{multideg } f$  and  $\beta = \text{multideg } g$ , consider

$$\gamma := (\gamma_1, \dots, \gamma_n), \quad \text{where } \gamma_i = \max\{\alpha_i, \beta_i\}.$$

The monomial  $x^\gamma$  is called the *least common multiple* of the monomials  $LM f$  and  $LM g$  and is denoted  $LCM(LM f, LM g) := x^\gamma$ . The expression

$$S(f, g) := \frac{x^\gamma}{LT f} \cdot f - \frac{x^\gamma}{LT g} \cdot g$$

is called the *S-polynomial* (also *syzygy*, or pair) of the polynomials  $f, g$ .

This is a tool for the elimination of leading terms. The Gaussian elimination is a special case of this procedure for polynomials of degree one. However, during the general procedure, it may happen that the degrees of the resulting polynomials are higher even though the original leading terms are removed.

For instance, consider the polynomials

$$f = x^3y^2 - x^2y^3 + x, \quad g = 3x^4y + y^2$$

of degree 5 in  $\mathbb{R}[x, y]$  with the  $<_{\text{grlex}}$  ordering. Then,  $\gamma = (4, 2)$  and

$$S(f, g) = \frac{x^4 y^2}{x^3 y^2} f - \frac{x^4 y^2}{3x^4 y} g = x f - \frac{1}{3} y g = -x^3 y^3 + x^2 - \frac{1}{3} y^3,$$

which is a polynomial of degree 6.

**Theorem.** Let  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  be an ideal. Then,  $G = \{g_1, \dots, g_t\}$  is a Gröbner basis of  $I$  if and only if for each  $i \neq j$ , the remainder of the division  $S(g_i, g_j)/G$  is zero.

**PROOF.** Begin with a technical lemma which describes which cancellations may occur when expressing polynomials in terms of generators. More precisely, they can always be expressed in terms of S-polynomials.

**Lemma.** Consider a polynomial  $f = \sum_{i=1}^t c_i x^{\alpha_i} g_i$ , where  $c_1, \dots, c_t \in \mathbb{K}$  and  $\alpha_i + \text{multideg } g_i = \delta$  for a fixed  $\delta$  whenever  $c_i \neq 0$ . If  $\text{multideg } f < \delta$ , then there are  $c_{jk} \in \mathbb{K}$  such that



$$\sum_{i=1}^t c_i x^{\alpha_i} g_i = \sum_{j,k=1}^t c_{jk} x^{\delta - \gamma_{jk}} S(g_j, g_k),$$

where  $x^{\gamma_{jk}} = \text{LCM}(\text{LM } g_j, \text{LM } g_k)$ , and the degree of each monomial  $x^{\delta - \gamma_{jk}} S(g_j, g_k)$  is less than  $\delta$ .

**PROOF.** Let  $d_i := \text{LC } g_i$  and  $p_i = x^{\alpha_i} g_i / d_i$ . Clearly,  $c_i d_i = \text{LC}(c_i x^{\alpha_i} g_i)$  and  $\text{LC } p_i = 1$ . Since  $\text{multideg}(c_i x^{\alpha_i} g_i) = \delta$  and  $\text{multideg } f < \delta$ , it follows that  $\sum_{i=1}^t c_i d_i = 0$ . Express  $f$  as a combination of S-polynomials:

$$\begin{aligned} f &= \sum_{i=1}^t c_i d_i p_i = c_1 d_1 (p_1 - p_2) + (c_1 d_1 + c_2 d_2) (p_2 - p_3) + \\ &+ \dots + (c_1 d_1 + \dots + c_{t-1} d_{t-1}) (p_{t-1} - p_t) + \\ &+ \underbrace{(c_1 d_1 + \dots + c_t d_t)}_0 p_t. \end{aligned}$$

Each difference  $p_j - p_k$  can be expressed in terms of S-polynomials

$$\begin{aligned} \frac{x^\delta}{d_j x^{\delta - \alpha_j}} g_j - \frac{x^\delta}{d_k x^{\delta - \alpha_k}} g_k &= x^{\delta - \gamma_{jk}} \left( \frac{x^{\gamma_{jk}}}{\text{LT } g_j} g_j - \frac{x^{\gamma_{jk}}}{\text{LT } g_k} g_k \right) = \\ &= x^{\delta - \gamma_{jk}} S(g_j, g_k) \end{aligned}$$

Now the individual coefficients  $c_{jk}$  can be derived easily from these equalities.  $\square$

Now follows the proof of the theorem. The “ $\implies$ ” implication follows directly from the corollary of subsection 12.6.11.

For the reverse implication, consider a non-zero polynomial  $f \in I$ . It must be shown that under the implication assumption,  $\text{LT } f \in \langle \text{LT } g_1, \dots, \text{LT } g_t \rangle$ . If it is known that the polynomial can be expressed as  $f = \sum_{i=1}^t h_i g_i$  with the property that

$$\text{multideg } f = \max\{\text{multideg}(h_i g_i)\},$$



then  $LT f$  is necessarily divisible by one of the leading terms  $LT g_i$ , which means that  $G$  is a Gröbner basis.

Denote  $m_i := \text{multideg}(h_i g_i)$ ,  $\delta := \max\{m_1, \dots, m_t\}$ . Clearly,  $\text{multideg } f \leq \delta$ . Let polynomials  $h_1, \dots, h_t$  be chosen so that  $\delta$  is as small as possible. Since this is a monomial ordering, which, in particular, is a well-ordering, such a  $\delta$  exists.

It is necessary to prove that  $\text{multideg } f = \delta$ . Write

$$\begin{aligned} f &= \sum_{m_i=\delta} h_i g_i + \sum_{m_i<\delta} h_i g_i = \\ &= \sum_{m_i=\delta} (LT h_i) g_i + \sum_{m_i=\delta} (h_i - LT h_i) g_i + \sum_{m_i<\delta} h_i g_i. \end{aligned}$$

Clearly, the degrees of all the terms of the second and third sums are less than  $\delta$ . Admitting that  $\text{multideg } f < \delta$ , it follows that

$$\text{multideg} \left( \sum_{m_i=\delta} (LT h_i) g_i \right) < \delta.$$

Denote  $c_i x^{\alpha_i} := LT h_i$  and apply the lemma:

$$\sum_{m_i=\delta} (LT h_i) g_i = \sum_{m_i=\delta} c_i x^{\alpha_i} g_i = \sum_{j,k} c_{jk} x^{\delta-\gamma_{jk}} S(g_j, g_k).$$

It follows from the assumption of the theorem and the multivariate division algorithm that

$$S(g_j, g_k) = \sum_{i=1}^t p_{ijk} g_i$$

and, moreover,  $\text{multideg}(p_{ijk} g_i) \leq \text{multideg } S(g_j, g_k)$ . Denote  $q_{ijk} := x^{\delta-\gamma_{jk}} p_{ijk}$ , to obtain

$$x^{\delta-\gamma_{jk}} S(g_j, g_k) = \sum_{i=1}^t q_{ijk} g_i.$$

By the second part of the lemma,

$$\text{multideg}(q_{ijk} g_i) \leq \text{multideg}(x^{\delta-\gamma_{jk}} S(g_j, g_k)) < \delta.$$

Substitution yields

$$\begin{aligned} \sum_{m_i=\delta} (LT h_i) g_i &= \sum_{j,k} c_{jk} \left( \sum_{i=1}^t q_{ijk} g_i \right) = \\ &= \sum_{i=1}^t \left( \sum_{j,k} c_{jk} q_{ijk} \right) g_i. \end{aligned}$$

At the same time,

$$\text{multideg} \left( \sum_{j,k} c_{jk} q_{ijk} g_i \right) < \delta \quad \text{for } i = 1, \dots, t.$$

Substitute this into the original equality, to get  $f$  expressed as a combination of  $g_1, \dots, g_t$ , where the degrees of all terms are less than  $\delta$ . This contradicts the minimality of  $\delta$ , and so  $\text{multideg } f = \delta$ , whence  $LT f \in \langle LT g_1, \dots, LT g_t \rangle$ . So  $G$  is a Gröbner basis.  $\square$

**12.6.13. A naive algorithm for Gröbner bases.** The theorem just proved provides an efficient method for deciding whether a given basis is a Gröbner basis. For example, consider  $I = \langle x + y, y - z \rangle$ . The only relevant  $S$ -polynomial is



$$S(x + y, y - z) = \frac{xy}{x}(x + y) - \frac{xy}{y}(y - z) = xz + y^2.$$

The division yields  $xz + y^2 = z(x + y) + y(y - z)$ , so it is a Gröbner basis.

The following algorithm utilizes just this method to find a Gröbner basis of an ideal that is generated by a given  $s$ -tuple of polynomials  $F = (f_1, \dots, f_s)$ .

- (1)  $G := F, G' := \emptyset$
- (2) while  $G \neq G'$ 
  - (a)  $G' := G$
  - (b)  $\forall p, q \in G' : p \neq q$  do
    - (i)  $s := \overline{S(p, q)}^{G'}$
    - (ii) if  $s \neq 0$ 

$$G := G \cup \{s\}$$

If the algorithm ever terminates, then  $G$  contains a Gröbner basis. Thus, it suffices to verify that it terminates. However, in each iteration of the inner loop (2), i.e. when a non-trivial remainder is added, either the monomial ideal generated by  $LT G$  extends or it remains unchanged. Consequently, there is a non-decreasing chain of (monomial) ideals  $I_1 = LT(F) \subseteq I_2 \subseteq \dots \subseteq I_k \subseteq \dots$ . Denoting  $I = \cup_{k=1}^{\infty} I_k$ , then  $I$  is an ideal, and by Hilbert's theorem, it is finitely generated. However, this means that all generators of  $I$  already lie in one of the  $I_k$ . Therefore, from this  $k$  onwards,  $I_k = I_{k+1} = \dots$ <sup>13</sup>

Clearly, the stabilization of this chain of monomial ideals of the leading terms is equivalent to termination of the algorithm.

This algorithm is far from ideal. There are quite trivial inputs for which it returns wild results. Moreover, the output basis directly depends on the input, so the outputs for the same ideal defined by different bases may vary.

**12.6.14. Reduction of bases.** In order to recognize the generators which are needed in a Gröbner basis, it suffices to follow the leading terms. The first step of the discussion is simply to remove all elements which are not needed in this sense.



**Lemma.** Let  $G$  be a Gröbner basis of an ideal  $I$  and  $p \in G$  such that  $LT p \in \langle LT(G \setminus \{p\}) \rangle$ . Then,  $G - \{p\}$  is also a Gröbner basis of  $I$ .

**PROOF.** From the definition of the Gröbner basis,  $\langle LT I \rangle = \langle LT G \rangle$ . But  $LT p \in \langle LT(G \setminus \{p\}) \rangle$ , so

<sup>13</sup>The condition of stabilization of every non-decreasing chain of ideals is called the "ascending chain condition". Rings which satisfy ACC are called Noetherian (in honour of Emy Noether). Hilbert's theorem can also be formulated as "a polynomial ring over a Noetherian ring is again Noetherian".

$\langle LT(G \setminus \{p\}) \rangle = \langle LT G \rangle$ . Hence the proposition follows immediately.  $\square$

**Definition.** A Gröbner basis  $G$  of an ideal  $I$  is said to be *minimal* if and only if  $LC p = 1$  and  $LT p \notin \langle LT(G - \{p\}) \rangle$  for all  $p \in G$ .

For instance, consider  $\mathbb{K}[x, y]$  and  $\langle_{\text{grlex}}, I = \langle f_1, f_2 \rangle = \langle x^3 - 2xy, x^2y - 2y^2 + x \rangle$ . The mentioned algorithm produces the following five polynomials  $F = (f_1, \dots, f_5)$ :

$$F = (x^3 - 2xy, x^2y - 2y^2 + x, -x^2, -2xy, -2y^2 + x).$$

Nevertheless,  $LT f_1 = x^3 = -x LT f_3$  and  $LT f_2 = -\frac{1}{2}x LT f_4$ , so neither  $f_1$  nor  $f_2$  is needed.

However, this reduction is still insufficient, since redundancy may occur at the level of individual terms of the basis elements. For example, for every  $a$ , the set  $\{x^2 + axy, xy, y^2 - 1/2x\}$  is a minimal Gröbner basis of the considered ideal.

That is why the following concept is introduced:

**REDUCED GRÖBNER BASIS**

Let  $G$  be a Gröbner basis of an ideal  $I$ . A polynomial  $g \in G$  is said to be *reduced* for the basis  $G$  if and only if none of its monomials lies in  $\langle LT(G \setminus \{g\}) \rangle$ .  $G$  is said to be *reduced* if and only if for all  $p \in G$ ,  $LC p = 1$  and  $p$  is reduced for  $G$ .

Clearly, every reduced Gröbner basis is minimal. Moreover:

**Proposition.** *If a polynomial  $g$  is reduced for a minimal Gröbner basis  $G$  of an ideal  $I$ , then it is reduced for every minimal Gröbner basis  $G'$  of  $I$  which contains it.*

**PROOF.** In order to arrive at a contradiction, let  $G = \{g_1, \dots, g_s\}$ ,  $G' = \{g'_1, \dots, g'_t\}$  be two minimal Gröbner bases. Choose a term  $m$  of a polynomial  $g$  where  $m \in \langle LT(G' - \{g\}) \rangle$  (i.e.  $g$  is not reduced for  $G'$ ). Then,  $m = a_1 LT g'_1 + \dots + a_t LT g'_t$  for appropriate polynomials  $a_1, \dots, a_t$ . Since both  $G$  and  $G'$  are Gröbner bases of the same ideal,  $\langle LT G \rangle = \langle LT G' \rangle$ , which means that each  $LT g'_i$  can be expressed as a multiple of one of the  $LT g_1, \dots, LT g_s$ , except  $LT g$  (since  $G'$  is minimal). Hence  $m \in \langle LT(G \setminus \{g\}) \rangle$ , which contradicts the assumption that  $g$  is reduced for  $G$ .  $\square$

Everything is now available to prove the main result about the existence and uniqueness of a reduced Gröbner basis. This is the main achievement of this part of the chapter on algebra. It allows for a straightforward algorithm to eliminate variables in systems of polynomial equations.



**12.6.15. Theorem.** *Let  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  be a non-zero ideal. Then, for every monomial ordering, there is a unique reduced Gröbner basis of  $I$ . Moreover, every Gröbner basis can be reduced algorithmically.*

**PROOF.** Assume that  $G$  is a Gröbner basis of an  $I$ . By the lemma of the previous subsection, it can be assumed that  $G$  is minimal. (The minimizing algorithm is clear: it suffices to

check for divisibility of leading monomials in any order and to omit redundant elements of the basis.)

Assume that a polynomial  $g \in G$  is not reduced. Then in the division  $g/(G \setminus \{g\})$ , the leading term of  $g$  stays in the remainder, since it is not divisible by anything (the basis is minimal). Therefore,  $LT(\overline{g}^{G \setminus \{g\}}) = LT g$ , since nothing else can be the leading term of the remainder. Now, denote

$$g' := \overline{g}^{G \setminus \{g\}}, \quad G' := (G \setminus \{g\}) \cup \{g'\}.$$

This new system of generators  $G'$  is again a minimal Gröbner basis of  $I$ , because  $\langle LT G' \rangle = \langle LT G \rangle$ . That is,  $\langle LT G' \rangle = \langle LT I \rangle$ . By properties of the multivariate division algorithm, the polynomial  $g'$  is reduced for  $G'$ . If a polynomial  $h \neq g$  is reduced for  $G$ , then it is also reduced for  $G'$  by the above proposition.

With every reduction of one of the elements, the total number of terms in all polynomials of the reduced Gröbner basis decreases. Therefore, the algorithm terminates as soon as each element is reduced. Hence there is an algorithm for the construction of the reduced Gröbner basis.

It remains to prove uniqueness. Let there be two reduced Gröbner bases  $G, \tilde{G}$  of a non-zero ideal  $I$ . Then  $\langle LT G \rangle = \langle LT I \rangle = \langle LT \tilde{G} \rangle$ . Since this ideal is monomial, Dickson's lemma can be applied.

Recalling the construction of the basis in the proof of Dickson's lemma, there exists a unique monomial basis of a monomial ideal such that the coefficients of its elements equal 1, and no element of the basis divides another one.

By the definition of minimality, both  $LT G$  and  $LT \tilde{G}$  must be such. This means that  $LT G = LT \tilde{G}$ . Consequently for each  $g \in G$ , there is a unique  $\tilde{g} \in \tilde{G}$  such that  $LT g = LT \tilde{g}$ .

$g - \tilde{g} \in I$ . Since  $G$  is a Gröbner basis,  $\overline{g - \tilde{g}}^G = 0$ . The terms  $LT g, LT \tilde{g}$  cancel out in  $g - \tilde{g}$ . Since both the bases are reduced, none of the remaining terms of  $g - \tilde{g}$  may be divisible by any of  $LT G = LT \tilde{G}$ . Therefore, it must be in the remainder, which means that

$$g - \tilde{g} = \overline{g - \tilde{g}}^G = 0.$$

This proves the uniqueness. □

**12.6.16. Remarks.** Several of the previous questions are now answered: It can be decided efficiently whether or not a given polynomial



lies in a given ideal by means of the multivariate division and the Gröbner basis.

Because of the reduced Gröbner bases, it can be decided whether or not two ideals coincide – they simply need to have the same reduced Gröbner basis.

This means that it can be decided whether or not a polynomial equation lies in the ideal generated by a given system. Moreover, it can be decided efficiently whether or not two given systems generate the same ideal of consequences.

The above algorithmic construction depends on an appropriate monomial ordering. The answers to the questions are, of course, independent of such ordering.

As mentioned at the beginning of this chapter, the technique of Gröbner bases is one of the fundamentals of computer algebra. Of course, this algorithm is usually implemented using various tricks to make it faster. One can use the reduction technique as early as when creating the Gröbner basis in the fundamental algorithm from subsection 12.6.13, etc.

In the literature, one may find miscellaneous variations for non-commutative algebraic objects (e.g. for formal manipulations with differential operators). The algorithm for finding a Gröbner basis can be viewed as a special case of the Knuth-Bendix algorithm for rewriting some rules. This solves the problem of word equivalence in monoids that are given by generators and a set of equalities.



Last, but not least, the technique of Gröbner bases can be used in a much more sophisticated way in commutative algebra. In the algorithm, the syzygies of all pairs of generators of a finite basis can be found sequentially. These syzygies are a basis of the submodule of all relations between  $k$  elements  $g_1, \dots, g_k$  of the basis, that is, subsets  $S$  in the space  $(\mathbb{K}[x_1, \dots, x_n])^k$ . The algorithm can be extended to these subsets, to find distinguished generators of all relations between generators. As long as there are some non-trivial relations the procedure can continue. It can be proved that after at most  $n$  such steps, there exist no non-trivial relations. The number of generators of relations in each step provides detailed information about the topological properties of the corresponding affine variety  $\mathfrak{V}(g_1, \dots, g_k)$ .

**12.6.17. Elimination of variables.** We finish this chapter by an application of the above algorithms.



Consider the ring  $\mathbb{K}[x_{p+1}, \dots, x_n]$  to be a subring of  $\mathbb{K}[x_1, \dots, x_n]$ . These are polynomials with no occurrence of the variables  $x_1, \dots, x_p$ . It is a subring, but not an ideal.

ELIMINATION IDEALS

Let  $I = \langle f_1, \dots, f_s \rangle \subseteq \mathbb{K}[x_1, \dots, x_n]$ . For  $p = 1, \dots, n$ , define

$$I_p := I \cap \mathbb{K}[x_{p+1}, \dots, x_n].$$

This set is called the *p-th elimination ideal*. Note that  $I_p$  is an ideal only in  $\mathbb{K}[x_{p+1}, \dots, x_n]$ .

On the level of polynomial equations,  $I_p$  contains all equations which are consequences of the system  $f_1 = 0, \dots, f_s = 0$  and which contain only the variables  $x_{p+1}, \dots, x_n$ .

**Theorem (Elimination theorem).** *Let  $I \subseteq \mathbb{K}[x_1, \dots, x_n]$  be an ideal and  $G = \{g_1, \dots, g_m\}$  a Gröbner basis of  $I$  with respect to  $<_{lex}$ . Let the variables be ordered as  $x_1 >_{lex} x_2 >_{lex} \dots$ . Then, for each  $p = 0, \dots, n$ ,  $G_p := G \cap \mathbb{K}[x_{p+1}, \dots, x_n]$  is a Gröbner basis for the ideal  $I_p$ .*

*If  $G$  is minimal or reduced, then  $G_p$  is again minimal or reduced, respectively.*

PROOF. Without loss of generality, assume that  $G_p = \{g_1, \dots, g_r\}$ . Since  $G \subseteq I$ , it follows that  $G_p \subseteq I_p$ . The inclusion  $\langle G_p \rangle \subseteq I_p$  is trivial.

It needs to be verified for each polynomial  $f \in I_p$  that

$$f = h_1 g_1 + \dots + h_r g_r.$$

To do this, perform multivariate division by the original Gröbner basis  $G$ . Since  $f \in I$ , it follows that  $\bar{f}^G = 0$ , i.e.

$$f = h_1 g_1 + \dots + h_r g_r + h_{r+1} g_{r+1} \dots + h_m g_m.$$

Each of the polynomials  $g_{r+1}, \dots, g_m$  must contain at least one of the variables  $x_1, \dots, x_p$ , otherwise it would lie in  $G_p$ . By the properties of the lexicographic ordering, this variable must also be contained in  $LT g_{r+1}, \dots, LT g_m$ .

Recall the individual steps of the algorithm for multivariate division and the fact that  $f$  contains no monomial with any of  $x_1, \dots, x_p$ . Then  $h_{r+1} = \dots = h_m = 0$ , thus verifying that  $f \in \langle G_p \rangle$ .

Not only the desired inclusion but also the fact that the division  $f/G$  on  $I_p$  gives the same result as  $f/G_p$  is proved. For  $1 \leq i < j \leq r$ , consider the  $S$ -polynomials  $S(g_i, g_j)$ .

$$\overline{S(g_i, g_j)}^{G_p} = \overline{S(g_i, g_j)}^G = 0,$$

so  $G_p$  is a Gröbner basis of  $I_p$ .

It is clear that the property, for the basis, of being either minimal or reduced, is preserved.  $\square$

The only property of the lexicographic ordering used in the proof is that if a variable occurs in the polynomial  $f$ , then it occurs in its leading term as well. However, this condition is much weaker than that of the lexicographic ordering. Therefore, in actual implementations, one may use any ordering with the mentioned property. This usually leads to more efficient computations, since the pure lexicographic ordering usually leads to an unpleasant increase of the polynomials' degrees.

**12.6.18. Back to parametrized varieties.** The above theorem suggests an algorithm for finding an implicit representation of a variety defined in terms of polynomial parametrization. Tools necessary for work with the smallest varieties that contain the points defined by parametrization, are not here available, so a detailed discussion is omitted.

When the parametrization of a variety is given by polynomial relations

$$x_1 = f_1(u_1, \dots, u_k), \dots, x_n = f_n(u_1, \dots, u_k),$$

the reduced Gröbner basis of the ideal

$$\langle x_1 - f_1, \dots, x_n - f_n \rangle$$

can be computed in the lexicographic ordering where  $u_i < x_j$  for all  $i, j$ . From this basis, the reduced Gröbner basis of the elimination ideal  $I_k$  is obtained. This is precisely the required ideal and its implicit representation.

It suffices to use an ordering which guarantees that each  $u_i$  is before each  $x_j$ , so that the computation of the Gröbner

basis would eliminate  $u_i$ ; otherwise the ordering may be arbitrary. There is a chance that there is a more efficient computation than with the pure lexicographic ordering.

When the parametrization is rational, i.e.

$$x_i = \frac{f_i(t_1, \dots, t_m)}{g_i(t_1, \dots, t_m)},$$

it is perhaps natural to think of substituting the ideal

$$\langle x_1 g_1 - f_1, \dots, x_n g_n - f_n \rangle$$

into the above theorem. However, the result of this is usually not good. For instance, consider

$$x = \frac{u^2}{v}, \quad y = \frac{v^2}{u}, \quad z = u.$$

Here,  $I = \langle vx - u^2, uy - v^2, z - u \rangle$ , and the elimination yields  $I_2 = \langle z(x^2y - z^3) \rangle$ . However, the correct result is  $\mathfrak{V}(x^2y - z^3)$ . The computation has added an entire plane.

The problem is that the entire variety of zero points of the denominators in the parametrizations of individual variables is included in  $W = \mathfrak{V}(g_1, \dots, g_n)$ . Instead, perceive the parametrization  $F$  as a mapping  $F : (\mathbb{K}^k - W) \rightarrow \mathbb{K}^n$ . For the implicit situation, use the ideal

$$\begin{aligned} I &= \langle g_1 x_1 - f_1, \dots, g_n x_n - f_n, 1 - g_1 \cdots g_n y \rangle \subseteq \\ &\subseteq \mathbb{K}[y, t_1, \dots, t_m, x_1, \dots, x_n], \end{aligned}$$

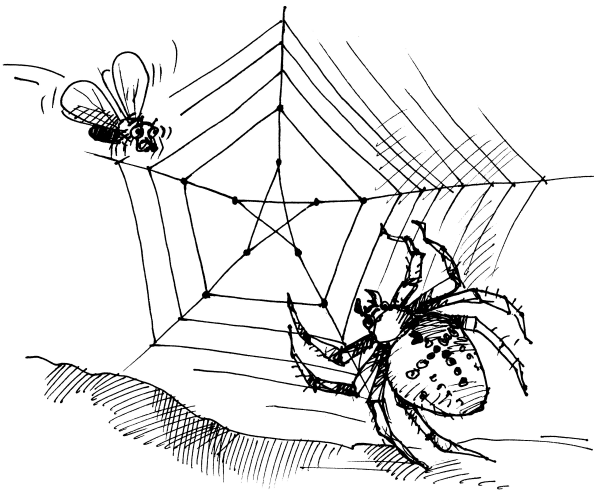
where the additional variable  $y$  enables avoidance of the zero spaces of the denominators. It can be shown that  $V(I_{k+1})$  is the minimal affine variety which contains  $F(\mathbb{K}^m - W)$ .

**Key to the exercises**



## Combinatorial methods, graphs, and algorithms

*Do we often prefer thinking in pictures?  
– yes, but we can compute discrete things only...*



### A. Fundamental concepts

One of the motives for creating graph theory was visualization of certain problems concerning relations. A human brain like thinking about entities it can imagine. Therefore, we like representing a binary relation with a graph whose vertices correspond to the elements and edges (lines between the elements) correspond to the fact that the given pair is related. Optionally, we can encode a relation in a more complicated way—use Hasse diagram (see 12.1.8), for instance. Partially ordered sets are almost always depicted this way. The relation of friendship or acquaintance between people can also be translated to graphs. This gives rise to a good deal of “relaxing” problems.

**13.A.1.** Prove that the number of odd-degree vertices in an undirected graph is always even.

**Solution.** Let  $G = (V, E)$  be an arbitrary undirected graph with degree  $d(v)$  at every  $v \in V$ . Adding the numbers of all degrees at the vertices we count all edges twice, and, therefore,  $\sum_{v \in V} d(v) = 2|E|$ . Let  $V_1 \subset V$  be the set of vertices

In this chapter, we return to problems concerning properties or mutual relations of (mainly) finite sets of objects. Combinatorial problems are already introduced in the second and third parts of chapter one.

Like number theory, combinatorics is a field of mathematics where the problems can often be formulated very easily. On the other hand, solutions can be much more difficult to find.

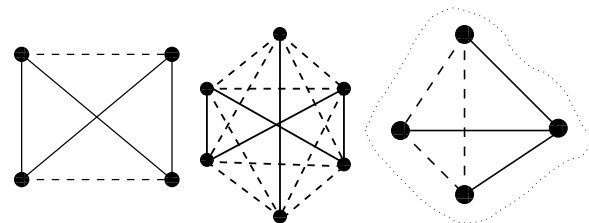
We begin with graph theory, and display a collection of useful algorithms based on this theory. At the end of the chapter methods of combinatorial computations are considered.

### 1. Elements of Graph theory

**13.1.1. Two examples.** Several people come to a party; some pairs of people know each other, while other people know nobody. (Acquaintance is assumed to be symmetric). How many people must be there in order to guarantee that there are either three people who all know each other, or there are three people with no mutual acquaintance?



Such situations can be aptly illustrated by a diagram. The points (or vertices) stand for the particular people of the party, the full lines represent pairs who know one another, while the dashed lines stand for pairs who do not know one another. Note that every pair of vertices is connected by either a full or a dashed line. The question is now reformulated as: how many vertices must be there in order that either there is a triangle whose sides are all full or a triangle whose sides are all dashed?



There is no such triangle in the left-hand diagram with four vertices. The example of a regular pentagon, in which all its outside edges are full, while all its diagonals are dashed (draw a picture!), shows that at least six vertices are required.

Such a triangle always exists if the number of vertices is at least six. To show this, consider a set of six vertices, each pair of which is joined by either a dashed line or a full line.

with all degrees and  $V_2 \subset V$ - set of vertices with even degrees. Then,

$$2|E| = \sum_{v \in V} d(v) = \sum_{v \in V_1} d(v) + \sum_{v \in V_2} d(v), \text{ and}$$

$$\sum_{v \in V_1} d(v) = 2|E| - \sum_{v \in V_2} d(v),$$

which is an even number. So,  $|V_1|$  is even. □

**13.A.2.** Does there exist a graph with degree sequence  $(3, 3, 2, 2, 2, 1)$ ?

**Solution.** In this sequence the number of odd-degree vertices equals odd number 3, which is impossible. □

**13.A.3.** Let  $G$  be a graph with minimum degree  $d > 1$ . Prove that  $G$  contains a cycle of length at least  $d + 1$ .

**Solution.** Let  $v_1 \dots v_k$  be a maximal path in  $G$ , i.e., a path that cannot be extended. Then any neighbour of  $v_1$  must be on the path, since otherwise we could extend it. Since  $v_1$  has at least  $d(G)$  neighbors, the set  $\{v_2, \dots, v_k\}$  must contain at least  $d(G)$  elements. Hence  $k \geq d(G) + 1$ , so the path has length at least  $d(G)$ . Now, neighbour of  $v_1$  that is furthest from  $v_1$  along the path must be  $v_i$  with  $i \geq d(G) + 1$ . Then  $v_1 \dots v_i v_1$  form a cycle of length at least  $d(G) + 1$ . □

**13.A.4.** Show that any graph with  $|V| \geq 2$  contains at least 2 vertices with equal degrees

**Solution.** Prove, using pigeon hole principle. Let  $|V| = n$ . Degree  $d(v)$  for any vertex  $v \in V$  can take values from 0 to  $n - 1$ . So, there are  $n$  possible distinct values for  $d(v)$ . However, if one of the degrees  $d(v) = n - 1$ , which means that vertex  $v$  is connected to any other vertex, no other degree can be 0. Thus,  $d(v)$  can take no more than  $n - 1$  distinct values, and, by pigeon hole principle, among  $n$  values of  $d(v)$  at least two coincide. □

**13.A.5.** Show that if  $n$  people attend a party and some shake hands with others, then at the end, there are at least two people who have shaken hands with the same number of people.

**Solution.** Let  $g$  be a graph with the set of vertices  $V$  being the set of people. If two people shake hands, that would represent an edge  $e$  between those two vertices. by previous problem, there will be at least two vertices with equal degrees. □

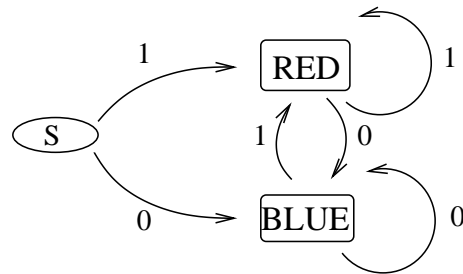
**13.A.6.** Show that if every connected component of a graph is bipartite, then the graph is bipartite.

**Solution.** Graph is bipartite, if we can divide its vertices into two subsets  $A$  and  $B$  such that every edge in the graph connects a vertex in set  $A$  to a vertex in set  $B$ . If the vertex sets of components are divided into sets  $A_i$  and  $B_i$ , we set  $A = \cup_i A_i$  and  $B = \cup_i B_i$ . Subsets  $A$  and  $B$  impose the graph with bipartite structure. □

If  $v$  is one of the vertices, then it is joined by five outgoing lines.

At least three of these lines are of one type, without loss of generality, full, joining to vertices  $v_A, v_B, v_C$ . Then either the triangle formed by the vertices  $v_A, v_B, v_C$ , contains only dashed lines, which is then the desired triangle, or one of its edges is full in which case there is a full triangle.

As another example, consider a black box which consumes one bit after another and shines in blue or in red according to whether the last bit is zero or one. Imagine this could be a light over the toilet door recognizing whether the last person came out (0) or in (1). Again, this scheme can be illustrated by a diagram:



The third vertex which has only two outgoing arrows represents the beginning of the system (before the first bit is sent).

Both situations share the same scheme: there is a finite set of objects represented by vertices. There is a set of their properties represented by connecting lines between particular vertices. The scheme can be modified by distinguishing the directions of the connecting lines by arrows.

Such a situation can be described in terms of relations; see the text from subsection 1.6.1 on in the sixth part of chapter one. But this is a complicated terminology for describing simple situations: In the first case, there is one set of people with two complementary symmetric and non-reflexive relations. In the second case, there are two antisymmetric relations on three elements.

**13.1.2. Fundamental concepts of graphs.** We use the terminology which corresponds to the latter diagrams.



#### GRAPHS AND DIRECTED GRAPHS

**Definition.** A *graph* (also an *undirected graph*) is a pair  $G = (V, E)$ , where  $V$  is the set of its *vertices* and  $E$  is a subset of the set  $\binom{V}{2}$  of all 2-element subsets of  $V$ .

The elements of  $E$  are called *edges* of the graph. The vertices of an edge  $e = \{v, w\}$ ,  $v \neq w$ , are called the *endpoints* of  $e$ . An endpoint of an edge is said to be *incident* to that edge. Two edges which share a vertex are called *adjacent*. Any two vertices which are the endpoints of an edge are called adjacent.

**13.A.7.** Show that a graph is bipartite if and only if it contains no cycles of odd length.

**Solution.** Suppose there are no cycles of odd length in the graph  $G$ . Choose any vertex from the graph and put it in set  $A$ . Follow every edge from that vertex and put all vertices at the other end in set  $B$ . Erase all the vertices that have been already used. Now for every vertex in  $B$ , follow all edges from each vertex in  $B$  and put the vertices on the other end in  $A$ , erasing all used vertices. Alternate back and forth in this manner until we cannot proceed. This happens when either we exhaust all the vertices or we encounter a vertex that is already in one set and needs to be moved to the other. If the latter occurs, that would represent an odd number of steps from it to itself, so there would be a cycle of odd length. If the graph is not connected, there may still be vertices that have not been assigned. Repeat same process until all vertices are assigned either to set  $A$  or to set  $B$ . Thus, graph  $G$  is bipartite. Suppose now that graph  $G$  is bipartite. Let  $c$  be a cycle  $v_{i_1} \dots v_{i_k}$  of length  $k$ . Vertices along this cycle alternate between subsets  $A$  and  $B$  of  $V$ . Also, the first  $v_{i_1}$  and the last  $v_{i_k}$  vertices on  $c$  should lie in the same set. Therefore, there is the same number of  $A$ -vertices and  $B$ -vertices on  $c$ , and, so,  $k$  is even.  $\square$

**13.A.8.** Show that any tree with at least two vertices is bipartite.

**Solution.** Since a tree does not contain any cycles, it also does not contain odd length cycles, and therefore, it is bipartite.  $\square$

**13.A.9.** Prove that if  $u$  is a vertex of odd degree in a graph, then there exists a path from  $u$  to some other vertex  $v$  of odd degree.

**Solution.** We build a path that does not reuse any edges. As we build the path, we erase edges we used in order not to use them again. Begin at vertex  $u$  and select an arbitrary path emanating from it. If, at any point, the path reaches a vertex of odd degree, we are done, but each time we arrive at a vertex  $t$  of even degree, we are guaranteed that there is another neighbouring vertex where we can move further. By passing through  $t$  we erase two edges from vertex  $t$ . Since the vertex originally was of even degree, coming in and going out reduces its degree by two, so it remains even. In this way, there is always a way to continue when we arrive at a vertex of even degree. Since there are only a finite number of edges, the tour must end eventually, and the only way it can end is if we arrive at a vertex of odd degree.  $\square$

**13.A.10.** If the distance  $d(u, v)$  between two vertices  $u$  and  $v$  that can be connected by a path in a graph is defined to be the length of the shortest path connecting them, then show that the distance function satisfies the triangle inequality:  $d(u, v) + d(v, w) \geq d(u, w)$ .

**Solution.** If one simply connects the paths from  $u$  to  $v$  to the path connecting  $v$  to  $w$  one obtains a path from  $u$  to  $w$  of length  $d(u, v) + d(v, w)$ . Since we are looking for the path

A *directed graph* is a pair  $G = (V, E)$ , where  $V$  is as above, but now,  $E \subseteq V \times V$ . The first of the vertices that define an edge  $e = (v, w)$  is called the *tail of the edge* and the other vertex is called its *head*. From the vertices' point of view,  $e$  is an *outgoing* edge of  $v$  and an *ingoing* edge of  $w$ . The directed edges are also called *arcs* or *arrows*. The head and the tail of a directed edge may be the same vertex; such an edge is called a *loop*.

Two directed edges are called *consecutive* if the tail of one of them is the head of the other one. Similarly, two vertices which are the head and the tail of an edge are called *consecutive*.

To every directed graph  $G = (V, E)$ , its *symmetrization* can be assigned. This is an undirected graph with the same set of vertices as  $G$ . It contains an edge  $e = \{v, w\}$  if and only if at least one of the edges  $e' = (v, w)$  and  $e'' = (w, v)$  belongs to  $E$ .

Graph theory provides an extraordinarily good language for thinking about procedures and deriving properties that concern finite sets of objects. They are a good example of a compromise between the tendency to "think in diagrams" and precise mathematical formulations.

The language of graph theory allows the adding of information about the vertices or edges in particular problems. For instance, the vertices of a graph can be "coloured" according to membership of the corresponding objects to several (pairwise disjoint) classes. Or the edges with several values can be labeled, and so on. The existence of an edge between differently coloured vertices can indicate a "conflict". For example, if the vertices are coloured red and blue according to membership to two groups of people with different interests and the edges represent adjacency at a dining table, then an edge connecting two differently coloured vertices can mean a potential conflict. Our first example from the previous subsection can thus be perceived as a graph with coloured edges. The statement we have checked there reads thus in the language of graph theory:

A graph  $K_n = (V, \binom{V}{2})$  with  $n \geq 6$  vertices and all possible edges which are labeled with two colours always contains a triangle whose sides are of the same colour.

The directed graph in the second example above, whose edges are labeled with zero or one, represents a simple *finite automaton*. This name reflects the idea that the graph describes a process which is, at any moment, in a state represented by the corresponding vertex. It changes to another state, in a step represented by one of the outgoing edges of that vertex. The theory of finite automata is not considered here.

**13.1.3. Examples of useful graphs.** The simplest case of graphs are those which contain no edges. There is no special notation for them. At the other extreme is a graph which contains all possible edges. This is called a *complete graph*, denoted by  $K_n$ , where  $n$  is the number of vertices of the graph. The graphs  $K_4$  and



of minimal length, if there is a shorter path it will be shorter than this one, so the triangle inequality will be satisfied.  $\square$

**13.A.11.** Show that in a directed graph where every vertex has the same number of incoming as outgoing edges there exists an Eulerian path for the graph.

**Solution.** Suppose a path starts at some vertex  $v$ . First edge on the path can be any edge emanating from  $v$ , that we delete from the set of possible further edges on a path. As following the path we arrive at any other vertex  $u$  we choose an edge that is still in the set of possible outgoing from  $u$  edges. We stop as we can no longer choose such an edge. This can happen only if we arrive again at  $v$  and all outgoing edges from  $v$  have been previously engaged in the path. Indeed, if this is any other vertex  $w$  then every arrival at  $w$  engages one incoming edge that can be followed by an outgoing edge as their number is the same. Hence, our path starts and ends at  $v$  and contains all edges of the graph.  $\square$

**13.A.12.** An  $n$ -cube  $Q_n$  is a cube in  $n$  dimensions that consists of vertices with coordinates 0 and 1 and edges, connecting neighbouring vertices, that may differ from each other in just one coordinate. Show that  $Q_n$  possesses a Hamiltonian circuit.

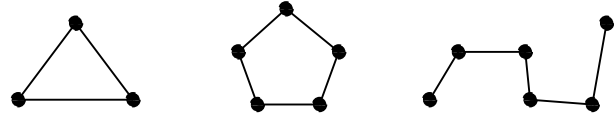
**Solution.** We prove by induction. The  $k + 1$ -dimensional cube  $Q_{k+1}$  can be considered as two copies of  $Q_k$  with every vertex of the first  $Q_k$  having first coordinate 0 and 1 for the second copy of  $Q_k$ . Consider some Hamiltonian cycle  $v_{i_1} \dots v_{i_k} v_{i_1}$  in  $Q_k$ . In the first copy of  $Q_k$  that we take a cycle  $0v_{i_1} \dots 0v_{i_k} 0v_{i_1}$ , and in the second copy  $1v_{i_1} \dots 1v_{i_k} 1v_{i_1}$ . Consider path  $0v_{i_1} \dots 0v_{i_k}, 1v_{i_k} 1v_{i_{k-1}} 1v_{i_1} 0v_{i_1}$  which forms a Hamiltonian cycle on  $Q_{k+1}$ .  $\square$

**13.A.13.** Show that a tree with  $n$  vertices has exactly  $n-1$  edges.

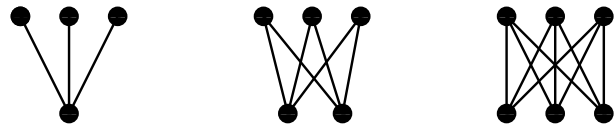
**Solution.** We proceed by induction. Suppose that every tree with  $k$  vertices has precisely  $k - 1$  edges. If the tree  $T$  contains  $k + 1$  vertices, we will show that it contains a vertex with a single edge connected to it. If not, start at any vertex, and start following edges marking each vertex as we pass it. If we ever come to a marked vertex, there is a loop in the edges which is impossible. But since each vertex is assumed to have more than one vertex coming out, there is never a reason that we have to stop at one, so we eventually encounter a marked vertex, which is a contradiction. Take the vertex with a single edge connecting to it, and delete it and its edge from the tree  $T$ . The new graph  $T_0$  will have  $k$  vertices. It must be connected, since the only thing we removed was a vertex that was not connected to anything else, and all other vertices must be connected. If there were no loops before, removing an edge certainly cannot produce a loop, so  $T_0$  is a tree. By the induction hypothesis,  $T_0$  has  $k-1$  edges. But to convert  $T_0$  to  $T$  we need to add one edge and one vertex, so  $T$  also satisfies the formula for the number of edges.  $\square$

$K_6$  are presented in the introductory subsection. The graph  $K_3$  is called a *triangle*.

An important type of graph, is a *path*. This is a graph whose vertices are ordered as  $(v_0, \dots, v_n)$  so that  $E = \{e_1, \dots, e_n\}$ , where  $e_i = \{v_{i-1}, v_i\}$  for all  $i = 1, \dots, n$ . A *path graph of length  $n$*  is denoted by  $P_n$ . If the first and last vertices coincide for the path graph ( $n \geq 3$ ), it is called a *cycle graph of length  $n$* , denoted by  $C_n$ . The graphs  $K_3 = C_3$ ,  $C_5$ , and  $P_5$  are shown in the following diagram.

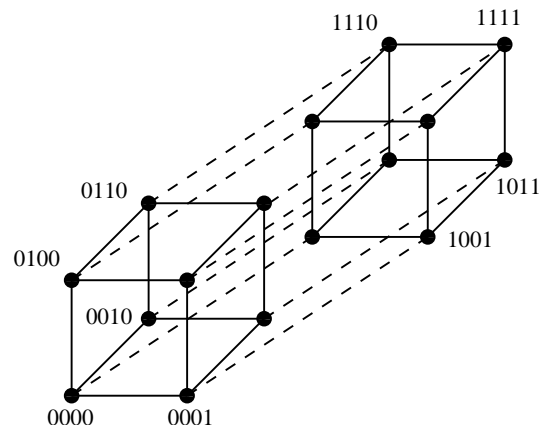


Another type of graph is the *complete bipartite graph*. Its vertices can be coloured with two (distinct) colours. All possible edges between vertices of different colours are present, but no other edges. Such a graph is denoted by  $K_{m,n}$ , where  $m$  and  $n$  are the numbers of vertices of particular colours. The diagram below illustrates the graphs  $K_{1,3}$ ,  $K_{2,3}$ , and  $K_{3,3}$ .



Another interesting example of a graph is the *hypercube  $H_n$*  in dimension  $n$ , whose vertices are the integers  $0, \dots, 2^n - 1$  and whose edges join those pairs of vertices whose binary expansions differ by exactly one bit. The following diagram depicts the hypercube  $H_4$ , with labels of the vertices indicated.

From the definition it follows that a hypercube of given dimension can always be composed from two hypercubes of dimension one lower, connecting them with edges in an appropriate way. These new edges between the two disjoint copies of  $H_3$  are the dashed ones in the diagram. Obviously, the hypercube  $H_4$  can be similarly decomposed in several ways (just looking at one fixed bit position, as is done with the very first position in the diagram).



Here are two more examples. The first is the *cycle ladder graph  $CL_n$*  with  $2n$  vertices. This consists of two cycle graphs  $C_n$  whose vertices are connected by edges according to their order in the cycles. The second is the *Petersen graph*.

**13.A.14.** If  $u$  and  $v$  are two vertices of a tree  $T$ , show that there is a unique path connecting them.

**Solution.** Since  $T$  is a tree, it is connected, and therefore there has to be at least one path connecting  $u$  and  $v$ . Suppose there are two different paths  $P$  and  $Q$  connecting  $u$  to  $v$ . Reverse  $Q$  to make a path  $Q'$  leading from  $v$  to  $u$ , and the path made by concatenating  $P$  and  $Q'$  leads from  $u$  back to itself. Now this path  $PQ'$  may not necessarily be a loop, since it may use some of the same edges in both directions, but we assumed that there are some differences between  $P$  and  $Q$ .

We can, from  $PQ'$ , generate a simple loop. Begin at  $u$  and continue back one node at a time until the paths  $P$  and  $Q'$  differ. At this point, the paths split. Continue along both paths of the beyond the bifurcation point until they join again for the first time, and this must occur eventually, since we know they are joined at the end. The two fragments of  $P$  and  $Q'$  form a smaller circuit in the tree, which is impossible, since  $T$  is a tree.  $\square$

**13.A.15.** If  $G$  is a connected graph and  $k \geq 2$  is the maximum path length, then any two paths in  $G$  with length  $k$  share at least one common vertex.

**Solution.** Suppose, not. Let  $P = v_{i_1} \dots v_{i_k}$  and  $Q = v_{j_1} \dots v_{j_k}$  be two paths of maximal length  $k$  that do not share any vertex. Vertices  $v_{i_1}$  and  $v_{j_1}$  are connected by a path  $R = u_{i_1} \dots u_{i_s}$ , where  $u_{i_1} = v_{i_1}$  and  $u_{i_s} = v_{j_1}$ . Path  $R$  coincides with  $P$  on some vertices, then it deviates from  $P$ . Maybe, returns to  $P$  again, then deviates from  $P$  for the last time and connects to  $Q$ . Denote this fragment of  $R$  by  $v_{i_r} = u_{i_m} u_{i_{m+1}} \dots u_{i_{m+t}} = v_{j_\ell}$ , where  $t \geq 1$ . Let  $P'$  be path  $v_{i_1} \dots v_{i_r}$ ,  $T' = u_{i_m} \dots u_{i_{m+t}}$  and  $Q'$  be the longest part of  $Q$  going forward or backwards from  $v_{j_\ell}$ , which length is obviously, at least,  $k + 1$ .  $\square$

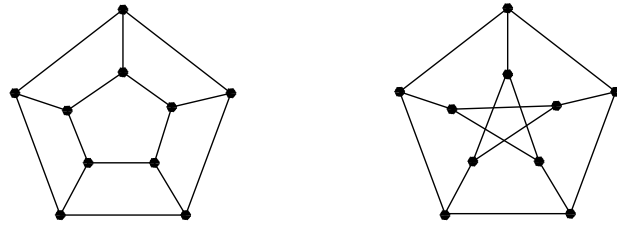
**13.A.16. Solution.**

$\square$

**13.A.17.** In a dormitory, there is a party held every night. Every time, the organizer of the party invites all of his/her acquaintances so that at the end of the party, all of the guests know each other. Suppose that each member of the dormitory has organized a party at least once, yet there are still two students who do not know each other. Show that they will not meet at the next party.

**Solution.** Consider the acquaintance graph of the students at the beginning (the vertices correspond to the students, and the edges to the acquaintances). We are going to show that if two students lie in the same connected component of this graph (i. e., there exists a chain of acquaintances beginning with one of the considered students and ending with the other one), see 13.1.10, then they will know each other as soon as each member of the dormitory has held a party. Indeed, consider the shortest path (acquaintance chain) between two students that lie in the same connected component. Every time someone from this path organizes a party, this path is made one shorter

This is somewhat similar to  $CL_5$ , yet it is actually the simplest counterexample for many propositions about graphs.



**13.1.4. Morphisms of graphs.** Mappings between the sets of vertices or edges which respect the considered structure are of great importance in graph theory. It is enough to consider mappings between the vertices only.



MORPHISMS OF GRAPHS

**Definition.** Let  $G = (V, E)$  and  $G' = (V', E')$  be two given graphs. A *morphism* (or *homomorphism*)  $f : G \rightarrow G'$  is a mapping  $f_V : V \rightarrow V'$  between the sets of vertices such that if  $e = \{v, w\}$  is an edge in  $E$ , then  $e' = \{f_V(v), f_V(w)\}$  is an edge in  $E'$ .

In practice, there is no need to distinguish between the morphism  $f$  and the mapping  $f_V$ .

The definition is the same for directed graphs, using ordered pairs  $e = (v, w)$  as edges.

In the case of undirected graphs, the definition implies that if  $f(v) = f(w)$  for distinct  $v, w \in V$ , then they are not connected by an edge. On the other hand, such an edge is admissible for directed graphs provided the common image of the vertices has a loop.

An important special case of a morphism of a graph  $G$  is one whose codomain is  $K_m$ . Such a morphism is equivalent to a labeling of the vertices of  $G$  with  $m$  colours (or any other names) of the vertices of  $K_m$  so that vertices of one colour are not adjacent. In this case, it is a (vertex) *colouring of the graph  $G$  with  $m$  colours*.

If a morphism  $f : G \rightarrow G'$  is a bijection between the sets of vertices such that the inverse mapping  $f^{-1}$  is also a morphism, then  $f$  is called an *isomorphism* of graphs. Two graphs are isomorphic if they differ only in the labeling of the vertices.

Every morphism of directed graphs is also a morphism of their symmetrizations. The converse is not true in general.

There are simple and extraordinarily useful examples of graph morphisms: namely a *path*, a *walk*, and a *cycle* in a graph:

(the organizer falls out). Since we assume that each of the students on the path has organized a party, the marginal students must know each other as well. Therefore, if there are two students who do not know each other even after everyone has held a party, then they lie in different connected components of the graph, so they will never meet at a party (in particular, not at the upcoming one).  $\square$

Now, we are going to practice the fundamental concepts of graph theory on simple combinatorial problems.

**13.A.18.** Determine the number of edges of each of the graphs  $K_6$ ,  $K_{5,6}$ ,  $C_8$ .

**Solution.** The complete graph  $K_6$  on 6 vertices has  $\binom{6}{2} = 15$  edges. The complete bipartite graph  $K_{5,6}$  (see 13.1.3) has  $5 \cdot 6 = 30$  edges. Finally, the cycle graph  $C_8$  has 8 edges.  $\square$

**13.A.19. Degree sequence.** Verify whether each of the following sequences is the degree sequence (see 13.1.7) of some graph. If so, draw one of the corresponding graphs.

- i) (1, 2, 3, 4, 5, 6, 7, 8, 9),
- ii) (1, 1, 1, 2, 2, 3, 4, 5, 5).

**Solution.** First of all, we should check the necessary condition from (1). In the former case, we have  $1 + \dots + 9 = \frac{1}{2} \cdot 9 \cdot 10 = 45$ , so the condition is not satisfied. Therefore, the first sequence does not correspond to any graph.

As for the latter sequence, the sum of the wanted degrees equals 24, so the necessary condition is satisfied. Now, we proceed by the Havel–Hakimi theorem from subsection 13.1.7.

$$\begin{aligned} (1, 1, 1, \underline{2}, \underline{2}, \underline{3}, \underline{4}, \underline{5}, \underline{5}) &\longleftrightarrow (1, 1, 1, \underline{1}, \underline{1}, \underline{2}, \underline{3}, \underline{4}) \longleftrightarrow \\ &\longleftrightarrow (1, 1, 1, 0, 0, 1, 2) \longleftrightarrow (0, 0, 1, 1, \underline{1}, \underline{1}, \underline{2}) \longleftrightarrow \\ &\longleftrightarrow (0, 0, 1, 1, 0, 0) \longleftrightarrow (0, 0, 0, 0, \underline{1}, \underline{1}) \longleftrightarrow \\ &\longleftrightarrow (0, 0, 0, 0, 0). \end{aligned}$$

Of course, it was not necessary to execute the procedure to the very end. We could have finished as soon as we saw that the obtained sequence indeed is the degree sequence of some graph. Now, we construct the corresponding graph “backwards” (however, we must take care to always add edges to vertices of appropriate degrees—it is this place where we have the option and can obtain non-isomorphic graphs with the same degree sequence). One of the possible outcomes is the following graph (the order in which each vertex was selected is written inside it):

WALKS, PATHS, TRAILS, AND CYCLES

A *walk of length  $n$*  in a graph  $G$  is a morphism  $s : P_n \rightarrow G$ . Both vertices and edges may repeat in the image.

A *trail* is a walk, where vertices are allowed to repeat, but edges are not allowed to repeat.

A *path of length  $n$*  in a graph  $G$  is any morphism  $p : P_n \rightarrow G$  such that  $p$  is an injective mapping. The images of the vertices  $v_0, \dots, v_n$  from  $P_n$  are pairwise distinct.

A *cycle of length  $n$*  in a graph  $G$  is any morphism  $c : C_n \rightarrow G$  such that  $c$  is an injective mapping of the vertices.

For simplicity, the morphism is often identified with its image. Walks are often written explicitly in the form  $(v_0, e_1, v_1, \dots, e_n, v_n)$ , where  $e_i = \{v_{i-1}, v_i\}$  for  $i = 1, \dots, n$ .

A walk can be thought of as the trajectory of a “pilgrim” moving from the vertex  $f(v_0)$  to the vertex  $f(v_n)$ , not stopping at any vertex of an (undirected) graph.  $P_n$  always contains an edge connecting the adjacent vertices  $v_{i-1}$  and  $v_i$ , while loops are not admitted in undirected graphs. The pilgrim can enter a vertex more than once or even go along an edge already visited. The pilgrim making a “trail” is a little wiser – he does not go along an edge already visited for the second time on his walk from the initial vertex  $f(v_0)$  to the terminal vertex  $f(v_n)$ .

**13.1.5. Subgraphs.** The images of paths, walks, and cycles are examples of *subgraphs*, but not in the same way.

SUBGRAPHS

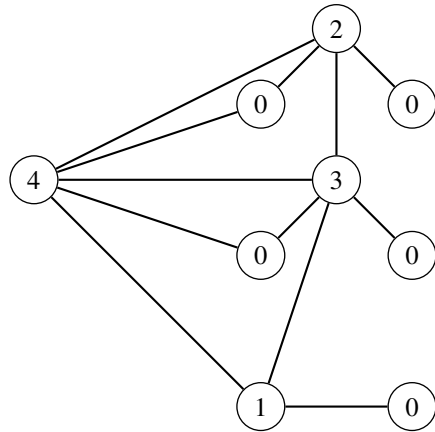
**Definition.** A graph  $G' = (V', E')$  is a subgraph of a graph  $G = (V, E)$  if and only if and only if  $V' \subseteq V$ , and  $E' \subseteq E$ .

Consider a graph  $G = (V, E)$ . Choose a subset  $V' \subseteq V$ . The largest subgraph (with respect to the number of edges) with  $V'$  as its set of vertices is called an *induced subgraph*. It is the graph  $G' = (V', E')$ , where an edge  $e \in E$  belongs to  $E'$  if and only if both of its endpoints lie in  $V'$ . Therefore, the set  $E'$  of  $G'$ 's edges is given as the intersection  $E \cap \binom{V'}{2}$ .

A *spanning subgraph* (also *factor*) of a graph  $G = (V, E)$  is any graph  $G' = (V, E')$  with  $V = V'$ . Hence  $G'$  has the same vertex set as  $G$ , but the set of edges  $E' \subset E$  may be arbitrary. A *clique* is a subgraph of the graph  $G$  which is isomorphic to a complete graph.

Every subgraph can be constructed by a step-by-step application of these two cases – first, select  $V' \subseteq V$ , then choose the target edge set  $E'$  in the subgraph induced on  $V'$ .

Every image of a homomorphism (vertices as well as edges) forms a subgraph.



□

**13.A.20.** Find the number of pairwise non-isomorphic complete bipartite graphs with 1001 edges.

**Solution.** A complete bipartite graph  $K_{m,n}$  has  $m \cdot n$  edges. Therefore, the problem can be stated as follows: In how many ways can we write the integer 1001 as the product of two integers? Since  $1001 = 7 \cdot 11 \cdot 13$ , we get  $1001 = 1 \cdot 1001 = 7 \cdot (11 \cdot 13) = 11 \cdot (7 \cdot 13) = 13 \cdot (7 \cdot 11)$ .

Thus, there are four non-isomorphic complete bipartite graphs having 1001 edges:

$$K_{1,1001}, K_{7,143}, K_{11,91} \text{ and } K_{13,77}. \quad \square$$

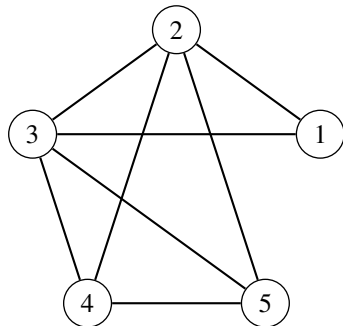
**13.A.21.** Find the number of graph homomorphisms (see 13.1.4)

- a) from  $P_2$  to  $K_5$ ,
- b) from  $K_3$  to  $K_5$ .

**Solution.** We can see from the definition of the graph homomorphism that the only condition which must be satisfied is that adjacent vertices must not be mapped to the same vertex.

- a)  $5 \cdot 4 \cdot 4 = 80$ . □
- b)  $5 \cdot 4 \cdot 3 = 60$ .

**13.A.22. Number of walks.** Using the adjacency matrix (see 13.1.8), find the number of trails of length 4 from vertex 1 to vertex 2 in the following graph:



**Solution.** The adjacency matrix of the given graph is

$$A_G = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

**13.1.6. How many non-isomorphic graphs are there?** It



is easy to draw all graphs (up to isomorphism) with a predetermined small number of vertices (three or four for instance). Generally this is a complicated combinatorial problem. It is often difficult to decide whether or not two given graphs are isomorphic.

**Remark.** This problem, known as the *Graph isomorphism problem*, is a somewhat peculiar member of the class  $\mathbf{NP}^1$  – it is known neither whether it is  $\mathbf{NP}$ -complete nor whether it can be solved in polynomial time. This is a special case of the problem of deciding whether or not a given graph is isomorphic to a subgraph of another graph. This *Subgraph isomorphism problem* is known to be  $\mathbf{NP}$ -complete.

Just to get some feeling, let us find a rough lower estimate of the total number of non-isomorphic graphs. There are the same number of graphs on a given set of  $n$  vertices as the number of subsets of the edge set. A  $k$ -element set has  $2^k$  subsets. There are at most  $n!$  graphs isomorphic to a given one, since this is the number of bijections between  $n$ -element sets. It follows that there are at least

$$k(n) = \frac{2^{\binom{n}{2}}}{n!}$$

pairwise non-isomorphic graphs. From this,

$$\log_2 k(n) = \binom{n}{2} - \log_2 n! \geq \frac{n^2}{2} \left( 1 - \frac{1}{n} - \frac{2 \log_2 n}{n} \right),$$

since  $n! \leq n^n$ . The asymptotic estimate for large  $n$ :

$$\log_2 k(n) \geq \frac{1}{2}n^2 - O(n \log_2 n),$$

follows. (See the notation for asymptotic bounds from subsection 6.1.12 on page 520). Thus the logarithm of the number of non-isomorphic graphs grows at least as fast as  $n^2$ .

**13.1.7. Vertex degree and degree sequence.** It is relatively



easy to verify that two given graphs are *not* isomorphic. Since isomorphic graphs differ in the relabeling of the vertices only, they share all numerical and other characteristics which are not changed by the relabeling. Simple data of this type includes, for instance, the number of edges incident to particular vertices.

<sup>1</sup>Wikipedia, *NP (complexity)*, [http://en.wikipedia.org/wiki/NP\\_\(complexity\)](http://en.wikipedia.org/wiki/NP_(complexity)) (as of Aug. 7, 2013, 13:44 GMT).

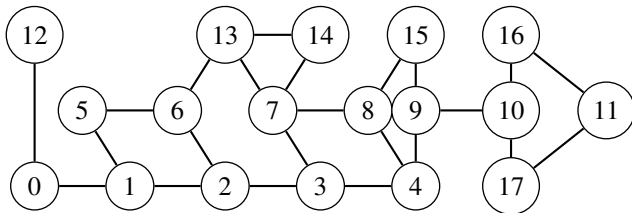
The number of walks of length 4 from vertex 1 to vertex 2 is the element at  $[1, 2]$  in the matrix  $A_G^4$ . Since

$$A_G^2 = \begin{pmatrix} 2 & 1 & 1 & 2 & 2 \\ 1 & 4 & 3 & 2 & 2 \\ 1 & 3 & 4 & 2 & 2 \\ 2 & 2 & 2 & 3 & 2 \\ 2 & 2 & 2 & 2 & 3 \end{pmatrix},$$

we have  $(A_G^4)_{1,2} = (2, 1, 1, 2, 2) \cdot (1, 4, 3, 2, 2)^T = 17$ . Therefore, there are 17 walks of length 4 between the vertices 1 and 2.  $\square$

**13.A.23.** A *cut edge* (also called bridge) in a graph is such an edge that its removal increases the number of connected components of the graph. Similarly, a *cut vertex* (also called articulation point) is a vertex with this property, i. e., when removed (with the edges incident to it, of course), the graph splits up into more connected components.

Find all cut edges and vertices of the following graph:



**13.A.24.** Prove that a Hamiltonian graph (see 13.1.13) must be 2-vertex-connected. Give an example of a graph which is 2-vertex-connected yet does not contain a Hamiltonian cycle.

**Solution.** Considering any pair of vertices in a Hamiltonian graph, there are two disjoint (except for the two vertices) paths between them (the “arcs” of the Hamiltonian cycle). Therefore, if we remove one of the vertices, the graph clearly remains connected (the vertex to be removed lies on one of the two paths only). As for the example of a non-Hamiltonian graph which is 2-vertex-connected, we can recall the Petersen graph (see the picture at the beginning of this chapter).  $\square$

**13.A.25.** Determine the number of cycles (see 13.1.3) in the graph  $K_5$ .

**Solution.** We sort the cycles by their lengths, i. e., we count separately the numbers of cycles upon three, four, and five vertices. A cycle of length three is determined uniquely by its three vertices, and there are  $\binom{5}{3}$  ways how to choose them. A cycle of length four is determined by its vertices (which can be chosen in  $\binom{5}{4}$  ways) and the pair of neighbors of a fixed vertex (the pair can be selected from the remaining three vertices in  $\binom{3}{2}$  ways). Finally, a cycle of length five is given by the pair of neighbors of a fixed vertex as well as the other neighbor (from the two remaining) of a fixed vertex of this

VERTEX DEGREE AND DEGREE SEQUENCE

The *degree* of a vertex  $v \in V$  in a graph  $G = (V, E)$  is the number of edges from  $E$  incident to  $v$ . It is denoted by  $\deg v$ .

The *degree sequence* of a graph  $G$  with vertices  $V = (v_1, \dots, v_n)$  is the sequence

$$(\deg v_1, \deg v_2, \dots, \deg v_n).$$

Sometimes, it is required that the sequence be sorted in ascending or descending order rather than correspond to the selected order of vertices.

In the case of directed graphs, distinguish between the *indegree*  $\deg_+ v$  of a vertex  $v$  and its *outdegree*  $\deg_- v$ . A directed graph is said to be *balanced* if and only if for all vertices  $v$   $\deg_- v = \deg_+ v$ .

The degree sequence of a graph (and its isomorphic copies) is unique up to permutation. Therefore, if the degree sequences of two graphs differ not merely by permutation, then the graphs are not isomorphic. The converse statement is not true in general. Two non-isomorphic graphs with the same degree sequence are the graph  $G = C_3 \cup C_3$  which has degree sequence  $(2, 2, 2, 2, 2, 2)$ , and  $C_6$ . However,  $C_6$  contains a path of length 5, but  $C_3 \cup C_3$  does not contain a path of length 5. Therefore, these two graphs cannot be isomorphic.

Since every edge has two endpoints, it is counted twice in the sum of the degree sequence (this condition is sometimes known as *handshaking lemma*). It follows that

$$(1) \quad \sum_{v \in V} \deg v = 2|E|.$$

In particular, the sum of the degree sequence must be even.

The following theorem<sup>2</sup> illustrates how useful invariants like the degree sequences can be. The proof is constructive. It describes an algorithm for constructing a graph with a given degree sequence if there is one, or shows that there is no such graph.



DECIDING ABOUT A GIVEN DEGREE SEQUENCE

**Theorem.** For any natural numbers  $0 \leq d_1 \leq \dots \leq d_n$ , there exists a graph  $G$  on  $n$  vertices with the above values as its degree sequence if and only if there exists a graph on  $n - 1$  vertices with degree sequence

$$(d_1, d_2, \dots, d_{n-d_n} - 1, d_{n-d_n+1} - 1, \dots, d_{n-1} - 1).$$

**PROOF.** If there exists a graph  $G'$  on  $n - 1$  vertices with degree sequence as stated in the theorem, then a new vertex  $v_n$  can be added to  $G'$ . Connect  $v_n$  with edges to the last  $d_n$  vertices of  $G'$ , thereby obtaining a graph  $G$  with the desired degree sequence.

<sup>2</sup>Proved independently by Václav J. Havel in 1955 in the Časopis pro pěstování matematiky (in Czech) and S. L. Hakimi in 1962 in the Journal of the Society for Industrial and Applied Mathematics.



pair. Altogether, there are

$$\binom{5}{3} + \binom{5}{4} \cdot \binom{3}{2} + \binom{5}{5} \cdot \binom{4}{2} \cdot \binom{2}{1} = 37$$

cycles. □

**13.A.26.** Determine the number of subgraphs (see 13.1.4) of the graph  $K_5$ .

**Solution.** Again, we count the number of subgraphs separately by the number  $v$  of their vertices:

- $v = 0$ . There is a unique graph on 0 vertices, the empty graph.
- $v = 1$ . There are 5 ways of selecting 1 vertex, resulting in 5 subgraphs.
- $v = 2$ . Two vertices can be chosen in  $\binom{5}{2}$  ways. Further, there may or may not be an edge between them. Altogether, we get  $\binom{5}{2} \cdot 2$  such subgraphs.
- $v = 3$ . Three vertices can be selected in  $\binom{5}{3}$  ways. For each pair of them, there may or may not be an edge. This results in  $\binom{5}{3} \cdot 2^{\binom{3}{2}}$  subgraphs.
- $v = 4$ . Here, we calculate  $\binom{5}{4} \cdot 2^{\binom{4}{2}}$  subgraphs.
- $v = 5$ . Finally, in this case, there are  $\binom{5}{5} \cdot 2^{\binom{5}{2}}$  subgraphs.

Altogether, we have found 1550 subgraphs of the graph  $K_5$ . □

**13.A.27.** Determine the number of paths between a fixed pair of different vertices in the graph  $K_7$ .

**Solution.** We sort the paths by their lengths. There is a unique path of length one (it consists of the edge that connects the selected vertices). There are five paths of length two (it may lead through any of the remaining vertices). There are  $5 \cdot 4$  paths of length three (we select the two vertices through which it leads, and their order matters). Similarly, there are  $5 \cdot 4 \cdot 3$  paths of length four,  $5 \cdot 4 \cdot 3 \cdot 2$  paths of length five, and also  $5!$  paths of length six. Clearly, there are no longer paths in  $K_7$ . Altogether, we have  $1 + 5 + 5 \cdot 4 + 5 \cdot 4 \cdot 3 + 5! + 5! = 326$  paths. □

At the end of this subsection, we present one more amusing problem.

**13.A.28.** The towns of a certain country are connected with roads. Each town is directly connected to exactly three other towns. Prove that there exists a town from which we can make a sightseeing tour such that the number of roads we use is not divisible by three.

**Solution.** First of all, we reformulate this problem in the language of graph theory. Our task is to prove that every 3-regular graph (i. e., such that the degree of every vertex equals three) contains a cycle whose length is not divisible by three. We will proceed by induction, and actually, we will prove a stronger proposition: every graph each of whose vertices has degree at least three contains a cycle whose length is not divisible by three. In fact, the original proposition could not be proved by induction since the induction hypothesis

The reverse implication is more difficult. The following needs to be proved. Suppose a fixed degree sequence  $(d_1, \dots, d_n)$  with  $0 \leq d_1 \leq \dots \leq d_n$  is given. Then there exists a graph whose vertex  $v_n$  is adjacent to exactly the last  $d_n$  vertices  $v_{n-d_n}, \dots, v_{n-1}$ .

The idea is simple – if any of the last  $d_n$  vertices  $v_k$  is not adjacent to  $v_n$ , then  $v_n$  must be adjacent to one of the prior vertices. We want to interchange the endpoints of two edges so that the vertices  $v_n$  and  $v_k$  become adjacent and the degree sequence remains unchanged.

Technically, this can be done as follows: Consider all graphs  $G$  with a given degree sequence and let, for each  $G$ ,  $\nu(G)$  denote the greatest index of a vertex which is not adjacent to the vertex  $v_n$ . Fix  $G$  to be such that  $\nu(G)$  is as small as possible. Then, either  $\nu(G) = n - d_n - 1$  (and the graph is obtained) or  $\nu(G) \geq n - d_n$ .

If the latter is true, then  $v_n$  is adjacent to one of the vertices  $v_i$ ,  $i < \nu(G)$ . Since,  $\deg v_{\nu(G)} \geq \deg v_i$ , there exists a vertex  $v_\ell$  which is adjacent to  $v_{\nu(G)}$ , but not to  $v_i$ . Replace the edge  $\{v_\ell, v_{\nu(G)}\}$  for  $\{v_\ell, v_i\}$  as well as  $\{v_i, v_n\}$  for  $\{v_{\nu(G)}, v_n\}$ , to get a graph  $G'$  with the same degree sequence, but with  $\nu(G') < \nu(G)$ , which contradicts the choice of  $G$ . (Draw a diagram!)

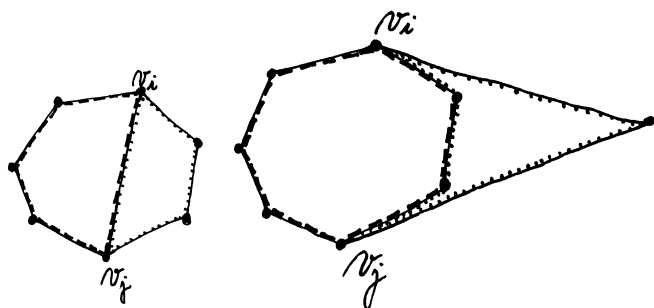
Therefore, the former possibility is true. So the graph is created by adding the last vertex and connecting it to the last  $d_n$  vertices with edges. □

The theorem describes an exact procedure for constructing a graph with a given degree sequence. If there is no such graph, the algorithm so indicates during the computation. Begin with the degree sequence in (say) ascending order. Then delete the largest value  $d$  and subtract one from  $d$  remaining values on the very right. Then sort the obtained degree sequence and continue with the above step until either there is an example of a graph with the current degree sequence or the degree sequence does not correspond to any graph. If, eventually, a graph is constructed after a number of steps, then one can reverse the procedure, adding one vertex in each step, connected to those vertices where ones were subtracted during the procedure. (Try examples by yourself!) The algorithm constructs only one of the many graphs which share the same degree sequence.

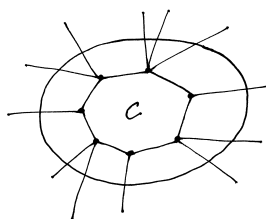
**13.1.8. Matrix representation.** The efficiency of graph representations is of importance for running algorithms. One of them is useful in theoretical considerations:



would be too weak. The induction will be carried on the number  $k$  of vertices of the graph. Clearly, the statement holds for  $k = 4$ . Now, consider a graph where the degree of each vertex is at least three and suppose that the statement is true for any such graph on fewer vertices. The reader should be able to prove that there exists a cycle in the graph. If its length is not divisible by three, we are done. Thus, suppose from now on that  $C = v_1v_2 \dots v_{3n}$ . Each vertex of this cycle is connected to at least one more (different from the neighbors on the cycle) in the graph. If there is a vertex  $v_i$  on the cycle that is connected to a vertex  $v_j$  on the cycle ( $j > i + 1$ ), then the lengths of the cycles  $v_1v_2 \dots v_iv_jv_{j+1} \dots v_{3n}$  and  $v_iv_{i+1} \dots v_j$  total up to  $3n + 2$ , so the length of at least one of them is not divisible by three, as wanted. The situation is similar if there are two vertices  $v_i$  and  $v_j$ ,  $1 \leq i < j \leq 3n$ , which are connected to the same vertex outside the cycle.



Therefore, suppose that each vertex of the cycle is connected to some vertices outside  $C$  and no two vertices of the cycle are adjacent to the same vertex outside. Then, we



can consider the graph which is obtained from the original one by replacing the vertices  $v_1, v_2, \dots, v_{3n}$  with a single vertex  $V$ .

In this new graph, there are also at least three edges leading from each vertex (including  $V$ ), so we can apply the induction hypothesis to it. Therefore, there is a cycle  $w_1w_2 \dots w_k$  where  $3 \nmid k$ . If it does not contain the vertex  $V$ , then it is a cycle in the original graph as well. If it does, then we proceed analogously as above: we consider two cycles whose lengths sum up to  $3n + 2k$ , so the length of at least one of them is not divisible by three. We have found the wanted cycle in every case, which finishes the proof.  $\square$

## B. Fundamental algorithms

Let us begin with breadth-first search and depth-first search, which serve as a basis for more sophisticated algorithms. Their actual implementations may differ; therefore, the answers to the following problems may be ambiguous.

**13.B.1.** Consider a graph on six vertices which are labeled  $1, 2, \dots, 6$ . A pair of vertices is connected with an edge if

## ADJACENCY MATRIX

The *adjacency matrix* of the (undirected) graph  $G = (V, E)$  is defined as follows. Fix a (total) ordering of its vertices  $V = (v_1, \dots, v_n)$ . Define the matrix  $A_G = (a_{ij})$  over  $\mathbb{Z}_2$  (entries are zero and ones)

$$a_{ij} = \begin{cases} 1 & \text{if the edge } e_{ij} = \{v_i, v_j\} \in E, \\ 0 & \text{if the edge } e_{ij} = \{v_i, v_j\} \notin E. \end{cases}$$

It is recommended to write explicitly the adjacency matrices of the graphs mentioned at the beginning of this chapter! By definition, adjacency matrices are symmetric.

There are straightforward generalizations of this concept for more general graphs. For oriented edges their directions may be indicated by the sign, multiple edges might be encoded by appropriate integers, etc.

If the matrix is stored in a two-dimensional array, then this method of graph representation is very inefficient. It consumes  $O(n^2)$  memory. However, if the graph is rather sparse, i.e. there are only a few edges, then almost all of the entries of the matrix are zeros. There are many methods of storing such matrices more efficiently.

The matrix representation of graphs is suggestive of linear algebra considerations. For example, there is the following beautiful theorem:

**Theorem.** Let  $G = (V, E)$  be a graph with vertices ordered as  $V = (v_1, \dots, v_n)$ , and let  $A_G$  be its adjacency matrix. Further, let  $A_G^k = (a_{ij}^{(k)})$  denote the entries of the  $k$ -th power of the matrix  $A_G = (a_{ij})$ .

Then,  $a_{ij}^{(k)}$  is the number of walks of length  $k$  between the vertices  $v_i$  and  $v_j$ .

**PROOF.** The proof is by induction on the length of the walks. For  $k = 1$ , the statement is simply a reformulation of the definition of the adjacency matrix. Suppose the proposition holds for a fixed positive integer  $k$ . Examine the number of walks of length  $k + 1$  between the vertices  $v_i$  and  $v_j$  for some fixed indices  $i$  and  $j$ . Each walk can be obtained by attaching an edge from  $v_i$  to a vertex  $v_\ell$  to a walk of length  $k$  between  $v_\ell$  and  $v_j$ . Further, each walk of length  $k + 1$  can be obtained uniquely in this way. Therefore, if  $a_{\ell j}^{(k)}$  denotes the number of walks of length  $k$  from  $v_\ell$  to  $v_j$ , then the number of walks of length  $k + 1$  is

$$a_{ij}^{(k+1)} = \sum_{\ell=1}^n a_{i\ell} \cdot a_{\ell j}^{(k)}.$$

This is exactly the formula for the product of the matrix  $A_G$  and the power  $A_G^k$ . It follows that the entries of the matrix  $A_G^{k+1}$  are the integers  $a_{ij}^{(k+1)}$ .  $\square$

**Corollary.** If  $G = (V, E)$  and  $A_G$  are as above, then each pair of vertices in  $G$  is connected by a path if and only if the matrix  $(A + \mathbb{E}_n)^{n-1}$  has only positive entries ( $\mathbb{E}_n$  is the  $n$ -by- $n$  identity matrix).

and only if the sum of their labels is odd. Describe the run of the breadth-first search algorithm on this graph. Which of the edges is visited at the end provided the search is initiated at vertex 5 and the neighbors of a given vertex are visited in ascending order?

**Solution.** The algorithm starts at vertex 5 and goes along the edges (5, 2), (5, 4), (5, 6), thereby visiting the vertices 2, 4, 6 (the queue of vertices to be processed is 2, 4, 6). The first vertex to have been visited is 2, so the algorithm continues the search from there, i. e., vertex 5 is processed and vertex 2 becomes active. The algorithm goes along the edges (2, 1), (2, 3), (2, 5) (the last one has already been used), thereby visiting the vertices 1 and 3 (the queue of vertices to be processed is 4, 6, 1, 3). Now, vertex 2 becomes processed and the first unprocessed vertex to have been visited becomes active. That is vertex 4. The algorithm discovers the edges (4, 1) and (4, 3), yet no new vertices. Vertex 4 becomes processed and vertex 6 becomes active. This leads to discovery of the edges (6, 1) and (6, 3). If the algorithm know the number of edges in the graph, it terminates at this moment. Otherwise, it goes through the vertices 1 and 3, finding out that there are no new edges or vertices, and then it terminates. In either case, the last edge to have been discovered is (3, 6).  $\square$

**13.B.2.** Consider a graph on six vertices which are labeled 1, 2, ..., 6. A pair of vertices is connected with an edge if and only if the sum of their labels is odd. Describe the run of the depth-first search algorithm on this graph. Which of the edges is visited at the end provided the search is initiated at vertex 5 and the neighbors of a given vertex are visited in ascending order?

**Solution.** The algorithm starts at vertex 5 and goes along the edges (5, 2), (5, 4), (5, 6), thereby visiting the vertices 2, 4, 6 in this order (the stack of vertices to be processed is 6, 4, 2). Vertex 5 becomes processed and the last vertex to have been visited (i. e., vertex 6) becomes active. The algorithm goes along the edges (6, 1) and (6, 3) (the edge (6, 5) has already been used), thereby visiting the vertices 1 and 3 (the stack of vertices to be processed is 3, 1, 4, 2). Now, vertex 2 becomes processed and the last unprocessed vertex to have been visited becomes active. This is vertex 3. The algorithm discovers the edges (3, 2) and (3, 4), so the stack becomes 4, 2, 1, 4, 2. Vertex 3 becomes processed and vertex 4 becomes active. This leads to discovery of the edge (4, 1), leaving the stack at 1, 2, 1, 2. The algorithm continues the search from vertex 1, visiting the last edge (1, 2). (Note: only unprocessed vertices are pushed into the stack.)  $\square$

**Remark.** If we had chosen the opposite edge priority, the edges would have been visited in the following order: (5, 2), (2, 1), (1, 4), (4, 3), (3, 2), (3, 6), (6, 1), (6, 5), (4, 5). Intuitively, the depth-first search can be perceived so that the algorithm examines the first undiscovered edge in each step.

**13.B.3.** Let the vertices of the graph  $K_6$  be labeled 1, 2, ..., 6. Write the order of edges of  $K_6$  in which they are visited by the depth-first search algorithm, supposing the search is initiated

**PROOF.**

$$(A + \mathbb{E}_n)^{n-1} = A^{n-1} + \binom{n-1}{1}A^{n-2} + \dots + \binom{n-1}{n-2}A + \mathbb{E}_n.$$

The entries of the resulting matrix are (using the notation as above)

$$a_{ij}^{(n-1)} + \dots + \binom{n-1}{\ell}a_{ij}^{(n-1-\ell)} + \dots + (n-1)a_{ij} + \delta_{ij},$$

where  $\delta_{ii} = 1$  for all  $i$ , and  $\delta_{ij} = 0$  for  $i \neq j$ .

This gives the sum of numbers of walks of length  $0, \dots, n-1$  between the vertices  $v_i$  and  $v_j$ , multiplied by positive constants. Therefore, it is non-zero if and only if there is a path between these vertices.  $\square$

**13.1.9. Remark.** Observe how permuting the vertices of  $V$  affects the adjacency matrix of the corresponding graph. It is not hard to see that each such permutation permutes both the rows and columns of the matrix  $A_G$  in the same way. Such a permutation can be given uniquely by the permutation matrix, each of whose rows and columns contain zeros only except for one entry, which is 1. If  $P$  is a permutation matrix, then the new adjacency matrix of the isomorphic graph  $G'$  is



$$A_{G'} = P \cdot A_G \cdot P^T,$$

where (the dots stand for matrix multiplication). The transposed matrix  $P^T$  is also the inverse matrix to  $P$ , since permutation matrices are orthogonal. Every permutation can be written as a composition of transpositions; hence every permutation matrix can be obtained as the product of the matrices corresponding to the transpositions.

Of course, this is exactly how the matrices of linear mappings change under the change of basis. Understanding the adjacency matrix as a linear mapping is often useful. For example, the adjacency matrix may be thought of as hitting vectors of zeros and ones (imagine the ones indicating active vertices of interest) and yielding vectors of integers (showing how many times the given vertices are arrived at from all active vectors along the edges in one step).

This observation also shows that the question whether two adjacency matrices describe isomorphic graphs is equivalent to asking for the equivalence of the matrices via a permutation matrix  $P$ .

**13.1.10. Connected components of a graph.** Every graph  $G = (V, E)$  naturally partitions into disjoint subgraphs  $G_i$  such that two vertices  $v \in G_i$  and  $w \in G_j$  are connected by a path if and only if  $i = j$ .

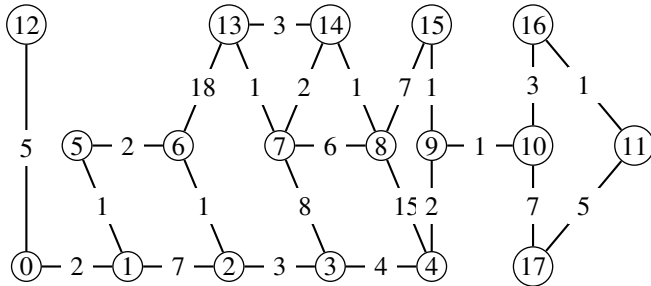


This procedure can be formalized as follows: Let  $G = (V, E)$  be an undirected graph. Define a relation  $\sim$  on the set  $V$ . Set  $v \sim w$  for vertices  $v, w \in V$  if and only if there exists a path from  $v$  to  $w$  in  $G$ . This relation is clearly a well-defined equivalence relation. Every class  $[v]$  of this equivalence determines the induced subgraph  $G_{[v]} \subseteq G$ , and the (disjoint) union of these subgraphs actually gives the original graph  $G$ . According to the definition of an equivalence

from vertex 3 and the neighbors of a given vertex are visited in ascending order. ○

**13.B.4.** Let the vertices of the graph  $K_6$  be labeled 1, 2, ..., 6. Write the order of edges of  $K_6$  in which they are visited by the breadth-first search algorithm, supposing the search is initiated from vertex 3 and the neighbors of a given vertex are visited in ascending order. ○

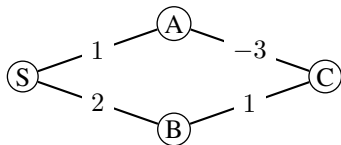
**13.B.5.** Apply Dijkstra's algorithm to find the shortest path from vertex number 9 to each of the other vertices.



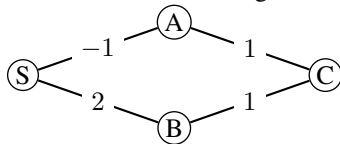
**13.B.6.** Give an example of

- i) a graph on at least 4 vertices which does not contain a negative cycle, yet Dijkstra's algorithm fails on it;
- ii) a graph on at least 4 vertices which contains a negative edge, yet Dijkstra's algorithm succeeds on it.

**Solution.** In both cases, we must be well aware how Dijkstra's algorithm works. Then, it is easy to find the wanted examples (apparently, there are many more possibilities). As for the first problem, we can consider the following graph (where  $S$  is the initial vertex):



If Dijkstra's algorithm is run from  $S$ , then it visits the vertex  $A$  and fixes its distance from  $S$  to 1. However, there is a shorter path, namely the path  $(S, B, C, A)$  of length 0. As for the second problem, consider the following:



**Bellman-Ford algorithm.** This algorithm is based on the same principle as Dijkstra's. However, instead of going through particular vertices, it processes them "simultaneously"—the *relaxation* loop (i. e., finding out whether the temporary distances of the vertices can be improved using a given edge) is iterated  $(|V|-1)$ -times over all edges. The advantage is that this approach works even with negative edges, and it

relation, no edge of the original graph can connect vertices that belong to different components. The subgraphs  $G_{[v]}$  are called *connected components* of the graph  $G$ .

A graph  $G = (V, E)$  is said to be *connected* if and only if it has exactly one connected component.

If the graph  $G$  is directed, then the definition is analogous to the case of undirected graphs – it is only required that there exist both paths from  $v$  to  $w$  and from  $w$  to  $v$  in order for the pair  $(v, w)$  to be related. Using this definition, strongly connected components can be discussed. On the other hand, it may only be required that the symmetrization of the graph be connected (in the undirected sense); then *weak connectedness* can be discussed.

**13.1.11. Multiply connected graphs.** It is useful to consider the concept of connectedness in a much stronger sense, i.e. to enforce a certain redundancy in the number of paths between vertices.

**Definition.** A graph  $G = (V, E)$  is said to be

- *k-vertex-connected* if and only if it has at least  $k + 1$  vertices, and remains connected whenever any  $k-1$  vertices are removed;
- *k-edge-connected* if and only if it has at least  $k$  edges, and remains connected whenever any  $k - 1$  edges are removed.

In the case  $k = 1$ , the definition simply says that the graph is connected (in both cases) since the condition is vacuously true. Stronger graph connectedness is desirable with any networks supporting some surfaces (roads, pipelines, internet connection, etc.) where the clients prefer considerable redundancy of the provided service for the case if several connections in the network (i.e. edges in a graph) or nodes in the network (vertices in a graph) break down.

In general, *Menger's theorem*<sup>3</sup> holds. It says that for every pair of vertices  $v$  and  $w$ , the number of pairwise edge-disjoint paths from  $v$  to  $w$  equals the minimum number of edges that must be removed so as to leave  $v$  and  $w$  in different components of the new graph. Similarly, the number of pairwise vertex-disjoint paths from  $v$  to  $w$  equals the number of vertices that must be removed in order to disconnect  $v$  from  $w$ .

We return to this topic in subsection 13.2.13. Right now, we consider the simplest interesting case in detail. These are graphs (on at least three vertices) such that deleting any one vertex does not destroy the connectedness.

**Theorem.** If  $G = (V, E)$  has at least three vertices, then the following conditions are equivalent:

- $G$  is 2-vertex-connected;
- every pair of vertices  $v$  and  $w$  in  $G$  lie on a common cycle;
- the graph  $G$  can be constructed from the triangle  $K_3$  by repeatedly adding and splitting edges.

<sup>3</sup>Karl Menger proved this as early as in 1927; that is, before graph theory came into being.

is able to detect negative cycles (if another iteration of the relaxation loop leads to a change, then there must be a negative cycle in the graph). However, we pay for that with increased time complexity.

**13.B.7.** Use the Bellman-Ford algorithm to find the shortest paths from the vertex  $S$  to all other vertices. Assume that the edges are ordered by the number of the tail (or head) and the initial vertex is the least one. Then, change the value of the edge  $(8, 6)$  from 18 to  $-18$ , execute the algorithm on this new graph, and show the detection of negative cycles.

picture skipped

**Solution.** According to the conditions, the edges are visited in the following order:  $(S,4)$ ,  $(S,7)$ ,  $(1,2)$ ,  $(1,5)$ ,  $(2,1)$ ,  $(2,3)$ ,  $(2,6)$ ,  $(3,7)$ ,  $(4,7)$ ,  $(4,8)$ ,  $(5,1)$ ,  $(5,6)$ ,  $(6,2)$ ,  $(6,5)$ ,  $(7,8)$ ,  $(8,6)$ . The vertex distances (potential higher values computed earlier during the same iteration are written in parentheses):

	$S$	1	2	3	4	5	6	7	8
1	0	$\infty$	$\infty$	$\infty$	1	$\infty$	22	3(6)	4
2	0	$\infty$	23	$\infty$	1	24	22	3	4
3	0	25(30)	23	26	1	24	22	3	3
4	0	25	23	26	1	24	22	3	3

Since the fourth iteration does not lead to any change, we can terminate the algorithm at this moment.

In the changed graph, the execution is as follows (for the sake of clarity, we do not write the values of vertices that are untouched by the change):

	$S$	1	2	3	4	5	6	7	8
1	0	$\infty$	$\infty$	$\infty$	1	$\infty$	-14	3(6)	4
2			-13			-12			
3		-11(-6)		-10			-19	-2	-1
4			-18			-17			
5		-16		-15			-24	-7	-6
6			-23			-22			
7		-21		-20			-29	-12	-11
8			-28			-27			
9		-26		-25			-34	-17	-16

The graph has 9 vertices, and since the ninth iteration changed the distance of one of the vertices, there is a negative cycle. Of course, we could have terminated the algorithm much earlier if we had noticed exactly what changes took place between the particular steps. Clearly, the values of the vertices 1, 2, 3, 5, 6, 7, 8 keep decreasing below all bounds. The algorithm can also be implemented so that it produces the tree of shortest paths and also finds the vertices lying on a negative cycle if there is one.  $\square$

**Paths between all pairs of vertices.** We often need to know the shortest paths between all pairs of vertices. Of course, we could apply the above algorithms to all initial vertices. However, there is a more effective method to do this. One of the possibilities is to use the similarity with matrix multiplication, which is the basis of the *Floyd-Warshall algorithm* (the best-known among algorithms of the *all pairs shortest paths* type), which:

- computes the distances between all pairs of vertices in time  $O(n^3)$ ;

**PROOF.** If the second proposition is true, there are at least two different paths between any two vertices. So deleting a vertex cannot destroy the connectedness and the first proposition follows.

Conversely, suppose the first proposition is true. Proceed by induction on the minimal length of a path between  $v$  and  $w$ . Suppose first that the vertices are the endpoints of an edge  $e$ , and that the shortest path is of length 1. If removing the edge  $e$  splits the graph into two components, then this would also occur if the vertex  $v$  is removed or if the vertex  $w$  is removed. Therefore, the graph is connected even without the edge  $e$ , so there is a path between  $v$  and  $w$ . This path, together with the edge  $e$ , forms a cycle.

For the induction hypothesis, assume that such a shared cycle is constructed for all pairs of vertices connected by a path whose length does not exceed  $k$ . Consider vertices  $v, w$  and one of the shortest paths between them:

$$(v = v_0, e_1, v_1, e_2, \dots, v_{k+1} = w)$$

of length  $k + 1$ . Then,  $v_1$  and  $w$  can be connected by a path of length at most  $k$ , hence they lie on a common cycle. Denote by  $P_1$  and  $P_2$  the corresponding two disjoint paths between  $v_1$  and  $w$ . Now, the graph  $G \setminus \{v_1\}$  is also connected, so there exists a path  $P$  from  $v$  to  $w$  which does not go through the vertex  $v_1$ , and this path must once meet either of the paths  $P_1, P_2$ . Without loss of generality, suppose that this occurs on the path  $P_1$ , at vertex  $z$ . Now, the cycle can be built: it consists of the part of  $P$  from  $v$  to  $z$ , the part of  $P_1$  from  $z$  to  $w$ , and  $P_2$  (directed the other way) from  $w$  to  $v$  (draw a diagram!). It follows that the second proposition is a consequence of the first proposition, and hence first condition is equivalent to the second one.

Suppose the third proposition is true. Neither splitting an edge nor adding a new one in a 2-vertex-connected graph destroys the 2-connectedness. So the first proposition follows from the third proposition.

It remains to prove that third proposition follows from the first proposition. From the first proposition,  $G$  is 2-connected, so there exists a cycle, which can be obtained from  $K_3$  by splitting edges. Consider the subgraph  $G' = (V', E')$  determined by this cycle, and consider an edge  $e = \{v, w\} \notin E'$  such that one of its endpoints lies in  $V'$ . If both of its endpoints lie there, a new edge can simply be added to the graph  $G'$ , which leads to the subgraph  $(V', E' \cup \{e\})$  in  $G$ , which contains more vertices and edges than the graph  $G'$ . Consider the remaining possibility, i.e.  $v \in V'$  while  $w \notin V'$ . Since  $G$  is 2-connected, it remains connected even if the vertex  $v$  is removed, and it contains a shortest path  $P$  between the vertex  $w$  and some vertex (denote it as  $v'$ ) in  $G'$  (apart from the removed vertex  $v$ ) and containing no other vertex from  $V'$ . Adding this path to the graph  $G'$ , together with the edge  $e$  (which can be done by adding the edge  $\{v, v'\}$  splitting it to the desired number of “new” vertices and edges), A new subgraph is obtained which satisfies the requirements

- starts with the matrix  $U_0 = A = (a_{ij})$  of edge lengths (setting  $u_{ii} = 0$  for each vertex  $i$ ) and then iteratively computes the matrices  $U_0, U_1, \dots, U_{|V|}$ , where  $u_k(i, j)$  is the length of the shortest path from  $i$  to  $j$  such that all of its inner vertices are among  $\{1, 2, \dots, k\}$ ;
- the matrices are computed using the formula

$$u_k(i, j) = \min\{u_{k-1}(i, j), u_{k-1}(i, k) + u_{k-1}(k, j)\}.$$

In other words, considering the shortest path from  $i$  to  $j$  which can go only through the vertices  $1, \dots, k$ , we can ask whether it uses the vertex  $k$ . If so, then this path consists of the shortest path from  $i$  to  $k$  and the shortest path from  $k$  to  $j$  (and these two paths use only the vertices  $1, \dots, k - 1$ ). Otherwise, the wanted path is also the shortest path from  $i$  to  $j$  which can go only through the vertices  $1, \dots, k - 1$ . Clearly, for  $k = |V|$ , we get the shortest paths between all pairs of vertices without any restrictions. Moreover, we can maintain the so-called predecessor matrix (i. e., the predecessor of each vertex on the shortest path from each vertex and update it as follows:

- Initialization:

$$(P_0)_{ij} = i \text{ for } i \neq j \text{ and } a_{ij} < \infty,$$

- In the  $k$ -th step, we update

$$(P_k)_{ij} = \begin{cases} (P_{k-1})_{kj}, & \text{if the path through } k \text{ is better,} \\ (P_{k-1})_{ij}, & \text{otherwise.} \end{cases}$$

As soon as the algorithm terminates, we can easily construct the shortest path between any pair of vertices  $u, v$ : we derive it from the matrix  $P = P_n = (p_{ij})$  (in the reverse order) as  $v, w = p_{uv}, p_{uw}, \dots$

**13.B.8.** Apply the Floyd–Warshall algorithm to the graph in the picture. Write the intermediate results into matrices. Show the detection of negative cycles. Maintain all information necessary for the construction of the shortest paths.

picture skipped

**Solution.** We proceed according to the algorithm, obtaining the following shortest-length matrices and predecessor matrices:

$$U_0 = \begin{pmatrix} 0 & 4 & -3 & \infty \\ -3 & 0 & -7 & \infty \\ \infty & 10 & 0 & 3 \\ 5 & 6 & 6 & 0 \end{pmatrix}, \quad P_0 = \begin{pmatrix} - & 1 & 1 & - \\ 2 & - & 2 & - \\ - & 3 & - & 3 \\ 4 & 4 & 4 & - \end{pmatrix};$$

$$U_1 = \begin{pmatrix} 0 & 4 & -3 & \infty \\ -3 & 0 & -7 & \infty \\ \infty & 10 & 0 & 3 \\ 5 & 6 & \mathbf{2} & 0 \end{pmatrix}, \quad P_1 = \begin{pmatrix} - & 1 & 1 & - \\ 2 & - & 2 & - \\ - & 3 & - & 3 \\ 4 & 4 & 1 & - \end{pmatrix};$$

$$U_2 = \begin{pmatrix} 0 & 4 & -3 & \infty \\ -3 & 0 & -7 & \infty \\ \mathbf{7} & 10 & 0 & 3 \\ \mathbf{3} & 6 & -1 & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} - & 1 & 1 & - \\ 2 & - & 2 & - \\ 2 & 3 & - & 3 \\ 2 & 4 & 2 & - \end{pmatrix};$$

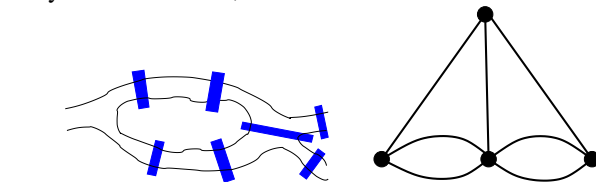
and contains more vertices than the considered graph  $G'$ . After a finite number of these steps, the entire graph  $G$  is built from the triangle  $K_3$ , as desired. The proof is complete.  $\square$

**13.1.12. Eulerian graphs.** There are problems of the type “draw a graph without removing the pencil from the paper”. In the language of graph theory, this can be stated as follows:

EULERIAN TRAILS

**Definition.** A trail which visits every edge exactly once and whose initial and terminal vertices are the same is called a *Eulerian trail*. Connected graphs that admit such a trail are called *Eulerian graphs*.

Of course, an Eulerian trail goes through every vertex at least once, but it can visit a vertex more than once. To draw a graph without removing the pencil from the paper while ending at the same point where one started means to find an Eulerian trail. The terminology refers to the classical story about the seven bridges in Königsberg. There, the task was to go for a walk and visit each of the bridges exactly once. The first proof that this is impossible is by Leonhard Euler, in 1736.



The situation is depicted in the diagram. On the left, there is a sketch of the river with the islands and bridges. The corresponding multigraph is caught in the right-hand diagram. The vertices of this graph correspond to the “connected land”, while the edges correspond to the bridges. If it is desired to do without the multiple edges (which have not been admitted so far), it would suffice to place another vertex inside each bridge (i.e. to split the edges with new vertices). Surprisingly, the general solution of this problem is quite simple, as shown by the following theorem. Of course, this also shows that Euler could not design the desired walk.

EULERIAN GRAPHS

**Theorem.** A graph  $G$  is Eulerian if and only if it is connected and all vertices of  $G$  have even degree.

**PROOF.** If a graph is Eulerian, for every vertex entered there is an exit. Therefore, the degree of every vertex is even. More formally: consider a trail that begins and ends at a vertex  $v_0$  and passes through all edges. Every vertex occurs once or more on this trail and its degree equals twice the number of its occurrences.

Now suppose that all vertices of a graph  $G$  have even degree. Consider the longest possible trail  $(v_0, e_1, \dots, v_k)$  in  $G$

$$U_3 = \begin{pmatrix} 0 & 4 & -3 & \mathbf{0} \\ -3 & 0 & -7 & -4 \\ 7 & 10 & 0 & 3 \\ 3 & 6 & -1 & 0 \end{pmatrix}, \quad P_3 = \begin{pmatrix} - & 1 & 1 & 3 \\ 2 & - & 2 & 3 \\ 2 & 3 & - & 3 \\ 2 & 4 & 2 & - \end{pmatrix};$$

$$U_4 = \begin{pmatrix} 0 & 4 & -3 & 0 \\ -3 & 0 & -7 & -4 \\ \mathbf{6} & \mathbf{9} & 0 & 3 \\ 3 & 6 & -1 & 0 \end{pmatrix}, \quad P_4 = \begin{pmatrix} - & 1 & 1 & 3 \\ 2 & - & 2 & 3 \\ 2 & 4 & - & 3 \\ 2 & 4 & 2 & - \end{pmatrix}.$$

Since there is no negative number on the diagonal of  $U_4$ , there is no negative cycle in the graph. Suppose we would like to find the shortest path from vertex 3 to vertex 1, for instance: The predecessor of 1 is  $P_4[3, 1] = 2$  and the predecessor of 2 is  $P_4[3, 2] = 4$ . Therefore, the wanted path is 3, 4, 2, 1 and its length is  $U_4[3, 1] = 6$ . □

**Hamiltonian graphs.** To decide whether a given graph is Hamiltonian is an NP-complete problem. Therefore, it might be useful to have some simpler necessary or sufficient conditions for this property at our disposal.

We mention three sufficient conditions: Dirac’s, Ore’s, and the Bondy–Chvátal theorem.

**Dirac:** Let a graph  $G$  with  $n \geq 3$  vertices be given. If each vertex of  $G$  has degree at least  $n/2$ , then  $G$  is Hamiltonian.

**Ore:** Let a graph  $G$  with  $n \geq 3$  vertices be given. If the sum of the degrees of each pair of non-adjacent vertices is at least  $n$ , then  $G$  is Hamiltonian.

The *closure* of a graph  $G$  is the graph  $cl(G)$  obtained from  $G$  by repeatedly adding an edge  $u, v$  such that  $u, v$  have not been adjacent and  $\deg(u) + \deg(v) \geq n$  until no such pair of vertices  $u, v$  exists.

**Bondy, Chvátal:** A graph  $G$  is Hamiltonian if and only if  $cl(G)$  is.

**13.B.9.** Prove the Bondy–Chvátal theorem.

**Solution.** Clearly, it suffices to prove that if  $G$  is Hamiltonian after addition of an edge  $\{u, v\}$  such that  $u, v$  have not been adjacent and  $\deg(u) + \deg(v) \geq n$ , then it is already Hamiltonian without this edge. Suppose that  $G + \{u, v\}$  is Hamiltonian, but  $G$  is not. Then, there exists a Hamiltonian path from  $u$  to  $v$  in  $G$ . It must hold for each vertex adjacent to  $u$  that its predecessor on this path is not adjacent to  $v$  (otherwise, there would be a Hamiltonian cycle in  $G$ ). Therefore,  $\deg(u) + \deg(v) \leq n - 1$ . □

**13.B.10.**

- i) Prove that the Bondy–Chvátal theorem implies Ore’s and Ore’s implies Dirac’s.
- ii) Give an example of a Hamiltonian graph which satisfies Ore’s condition but not Dirac’s.
- iii) Give an example of a Hamiltonian graph whose closure is not a complete graph.

**Solution.**

where no edge occurs twice or more. First, suppose for a moment that  $v_k \neq v_0$ . This would mean that the number of edges of the trail that enter or leave the vertex  $v_0$  is odd, so there must be an edge which is incident to  $v_0$  and not contained in the trail. However, then the trail can be prolonged while still using every edge of the graph at most once, which is a contradiction. Therefore,  $v_0 = v_k$ .

Define a subgraph  $G' = (V', E')$  of  $G$  as follows: It contains the vertices and edges of our fixed trail and nothing else. If  $V' \neq V$ , then (since the graph  $G$  is connected) there exists an edge  $e = \{v, w\}$  such that  $v \in V'$  and  $w \notin V'$ . However, then the trail can be “rotated” so that it begins and ends at the vertex  $v$ . It can be prolonged with the edge  $e$ , which contradicts the assumption of the greatest possible length. Therefore,  $V' = V$ .

It remains to show that  $E' = E$ . So suppose there is an edge  $e = \{v, w\} \notin E'$ . As above, the trail can be rotated so that it begins and ends at the vertex  $v$  and then it can continue the edge  $e$  – a contradiction. □

**Corollary.** A connected graph can be drawn without removing the pencil from the paper if and only if there are either no vertices of odd degree or exactly two of them.

**PROOF.** Let  $G$  be a connected graph with exactly two odd-degree vertices. Construct a new graph  $G'$  by attaching a new vertex  $w$  to the original graph  $G$  and connecting it to both the odd-degree vertices. This graph is Eulerian, and the Eulerian trail in it leads to the desired result.

On the other hand, if a graph  $G$  can be drawn in the desired way, then the graph  $G'$  is necessarily Eulerian, so the degrees of the vertices in  $G$  are as stated. □

The situation for directed graphs is similar. A directed graph is called *balanced* if and only if the outcoming and incoming degrees coincide, i.e.  $\deg_+(v) = \deg_-(v)$ , for all vertices  $v$ .

**Proposition.** A directed graph  $G$  is Eulerian if and only if it is balanced and its symmetrization is connected (i.e. the graph  $G$  is weakly connected).

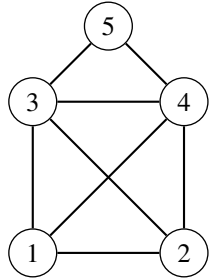
**PROOF.** The proof is analogous to the undirected case. (Work out the details yourself!) □

**13.1.13. Hamiltonian cycles.** Find a walk or cycle that visits every vertex of a graph  $G$  exactly once. Of necessity, such a walk can visit every edge at most once. Such a cycle is called a *Hamiltonian cycle* in the graph  $G$ . A graph is called *Hamiltonian* if and only if it contains a Hamiltonian cycle. This problem seems to be very similar to the above one of visiting every edge exactly once. But while the problem of deciding about the existence of a Eulerian trail is trivial, the problem of deciding whether a graph is Hamiltonian is **NP**-complete.

Of course, this problem can be solved by “brute force”. Given a graph on  $n$  vertices, generate all  $n!$  possible orders of the  $n$  vertices, and for each of them, verify whether it is a cycle in  $G$ .



- i) If a graph  $G$  satisfies Ore's condition, then its closure is a complete graph, which is Hamiltonian, of course. By the Bondy–Chvátal theorem, the original graph is Hamiltonian as well. Further if  $G$  satisfies Dirac's condition, then it clearly satisfies Ore's as well and thus is Hamiltonian.
- ii) Consider the following example:

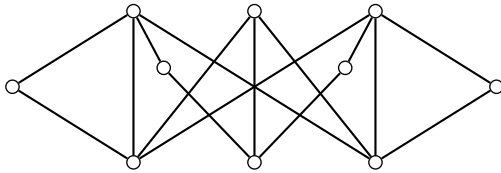


The degree of vertex 5 is 2, which is less than  $\frac{5}{2}$ . The sum of the degrees of any pair of (not only non-adjacent) vertices is at least 5.

- iii) The wanted conditions are satisfied by the cycle graphs  $C_n, n > 4$ , for which  $cl(C_n) = C_n$ .  $\square$

**Planar graphs.**

- 13.B.11.** Decide whether the graph in the picture is planar.



**Solution.** By the Kuratowski theorem (see page 1162), this graph is not planar since one of its subgraphs is a subdivision of  $K_{3,3}$ .  $\square$

- 13.B.12.** Decide whether there is a graph with degree sequence

$$(6, 6, 6, 7, 7, 7, 7, 8, 8, 8).$$

If so, is there a planar graph with this degree sequence?  $\bigcirc$

- 13.B.13.** What is the minimum number of edges of a hexahedron?

**Solution.** In any polyhedron, every face is bounded by at least three edges. At the same time, every edge belongs to two faces. If  $f$  is the number of faces and  $e$  the number of edges of the polyhedron, then we have  $3f \leq 2e$  (see also 13.1.20). For a hexahedron, this bound yields  $18 \leq 2e$ , i. e.,  $e \geq 9$ . Indeed, there exists a hexahedron with nine edges. It can be obtained by “gluing” two identical regular tetrahedra together along one face. Therefore, the minimum number of edges of a hexahedron is nine.  $\square$

This problem forms a vital field of research. For instance, in 2010, A. Björklund published a randomized algorithm based on the Monte Carlo method, which counts the number of Hamiltonian cycles in a graph on  $n$  vertices in time  $O(1,567^n)$ .<sup>4</sup>

Finding Hamiltonian cycles is desired in many problems related to logistics. For example, finding optimal paths for goods delivery.

**13.1.14. Trees.** Sometimes it is desired to minimize the number of edges in the graph while keeping it connected. Of course, this is possible if and only if there is at least one cycle in the graph. The graphs without cycles are extremely important as we shall see below.



FORESTS, TREES, LEAVES

A connected graph which does not contain a cycle is called a *tree*. A graph which does not contain a cycle is called a *forest* (a forest is not required to be connected). Every vertex of degree one in any graph is called a *leaf*.

The definition suggests an easily memorable “theorem”: *A tree is a connected forest.*

**Lemma.** Every tree with at least two vertices contains at least two leaves.

For any graph  $G$  with a leaf  $v$ , the following propositions are equivalent:

- $G$  is a tree;
- $G \setminus v$  is a tree.

**PROOF.** Let  $P = (v_0, \dots, v_k)$  be (any) longest possible path in a tree  $G$ . If the vertex  $v_0$  is not a leaf, then there is an edge  $e$  incident to it whose other endpoint  $v$  is not in  $P$  since this would form a cycle in the tree. Then the path  $P$  with this edge could be prolonged, which contradicts “longest”. So the vertex  $v_0$  is a leaf. The proof for the vertex  $v_k$  is similar. Thus, if the longest path is not trivial, then it must contain two leaves  $v_0$  and  $v_k$ .

Next, consider any leaf  $v$  of a tree  $G$ . Consider any two other vertices  $w, z$  in  $G$ . There exists a path between them, and no vertex on this path has degree one. Therefore, this path remains the same in  $G \setminus v$ . Hence the graph remains connected even after removing the vertex  $v$ . There is no cycle, since it is constructed by removing a vertex from a tree.

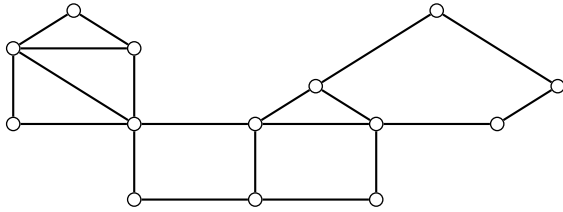
Conversely, if  $G \setminus v$  is a tree, then adding a vertex with degree 1 cannot create a cycle. The resulting graph is evidently connected.  $\square$

Trees can be characterized by many equivalent and useful properties. Some of them appear in the following theorem which is more difficult to formulate than to prove.

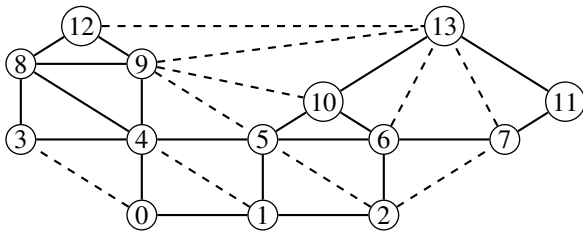
<sup>4</sup>Björklund, Andreas (2010), "Determinant sums for undirected Hamiltonicity", Proc. 51st Annual Symposium on Foundations of Computer Science (FOCS '10), pp. 173-182, arXiv:1008.0541, doi:10.1109/FOCS.2010.24.



**13.B.14.** Decide whether the given planar graph is maximal. Add as many edges as possible while keeping the graph planar.



**Solution.** The graph has 14 vertices and 20 edges, hence  $3|V| - 6 - |E| = 16$ . Therefore, it is not maximal, and 16 edges can be added so that it is still planar.



Ten (dashed) have been added. For the sake of clarity, the other 6 edges that connect the vertices of the “outer” 9-gon are not drawn.

□

**13.B.15.** Prove or disprove each of the following propositions.

- i) Every graph with fewer than 9 edges is planar.
- ii) Every graph which is not planar is not Hamiltonian.
- iii) Every graph which is not planar is Hamiltonian.
- iv) Every graph which is not planar is not Eulerian (see 13.1.12).
- v) Every graph which is not planar is Eulerian.
- vi) Every Hamiltonian graph is planar.
- vii) No Hamiltonian graph is planar.
- viii) Every Eulerian graph is planar.
- ix) No Eulerian graph is planar.

**Trees.**

**13.B.16.** Determine the code of the following graph as a

- i) plane tree,
- ii) tree.

**13.1.15. Theorem.** Let  $G = (V, E)$  be a graph. The following conditions are equivalent:

- (1)  $G$  is a tree.
- (2) For every pair of vertices  $v, w$ , there is exactly one path from  $v$  to  $w$ .
- (3)  $G$  is connected but ceases to be such if any edge is removed.
- (4)  $G$  does not contain a cycle, but the addition of any edge creates one.
- (5)  $G$  is a connected graph, and the number of its vertices and edges satisfies

$$|V| = |E| + 1.$$

**PROOF.** The properties 2–5 are satisfied in every tree. Indeed, by the previous lemma, every tree which has at least two vertices has a leaf  $v$ . It continues to be a tree when this leaf  $v$  is removed. Therefore, it suffices to show that if any of the statements 2–5 is true for a given tree, then it holds when a leaf is added to the tree as well. This is clear.

In the case of properties 2 and 3, the graph is connected, and their formulation directly excludes the existence of cycles. As for the fourth property, it suffices to verify that  $G$  is connected. However, any two of vertices  $v, w$  in  $G$  are either connected with an edge, or adding this edge to the graph creates a cycle. So there exists a path between them even without this edge.

The last implication can be proved by induction on the number of vertices. Suppose that all connected graphs on  $n$  vertices and  $n - 1$  edges are trees. The sum of vertex degrees of any graph on  $n + 1$  vertices and  $n$  edges is  $2n$ , so the graph must contain a leaf. It follows from the induction hypothesis that this graph can be constructed by attaching a leaf to a tree; hence it is also a tree. □

**13.1.16. Rooted trees, binary trees, and heaps.** Trees are often suitable structures for data storage. They permit basic database operations (e.g. finding a particular piece of information) efficiently.

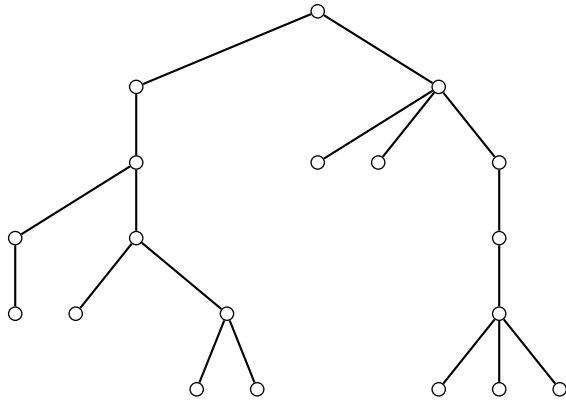


Since there is no cycle in a tree, fixing one vertex  $v_r$  defines the orientation of all edges. For every vertex  $v$ , there is exactly one path from  $v_r$  to  $v$ , so the orientation can be defined accordingly. Since there are no cycles, it is impossible for two such paths to force both orientations of a particular edge.

If one of the vertices of a tree is fixed, the situation is similar to a real tree in nature – there is a distinctive vertex which “grows from the ground”. Trees with a fixed distinguished vertex  $v_r$  are called *rooted trees*, and  $v_r$  is said to be the *root of the tree*.

In a rooted tree, the terms *successor* and *predecessor* of a vertex are defined as follows: a vertex  $w$  is a successor of  $v$  (or  $v$  is a predecessor of  $w$ ) if and only if the path from the root of the tree to the vertex  $w$  goes through  $v$  and  $v \neq w$ . If the vertices are directly connected with an edge, we can talk about a *direct successor* and a *direct predecessor*. More

○



**Solution.**

- i) Using the procedure from 13.1.18, we get the following code of the plane tree:

0 0 0001100100101111 1 0 0101000010101111 1 1

The highlighted vertex in the graph is indeed the appropriate candidate to be the root since it is the only element of the center of the tree.

- ii) As for the unique construction of a plane tree, we sort the descendants lexicographically in ascending order. Thus, the wanted code is

0000001010111101011 0000010110110011111. □

**13.B.17.** For each of the following codes, decide whether there exists a tree with this code. If so, draw the corresponding tree.

- 00011001111001,
- 0000011001001011111001010000101011111.

**Huffman coding.** We are working with plane binary trees where every edge is *colored* with a symbol of an output alphabet  $A$  (we often have  $A = \{0, 1\}$ ). The codewords  $C$  are those words over the alphabet  $A$  to which we translate the symbols of the input alphabet. Our task is to represent a given text using suitable codewords over the output alphabet.

We can easily see that it makes sense to require that the set of codewords be *prefix*, (i. e., no codeword can be a prefix of another one); otherwise, we could get into trouble when decoding.

We will use binary trees for the construction of binary prefix codes (i. e. over the alphabet  $A = \{0, 1\}$ ). We label the edges going from each vertex by 0 and 1. Further, we label the leaves of our tree with symbols of the input alphabet. This results in a prefix code over  $A$  for these symbols by concatenating the edge labels along the path from the root to the corresponding leaf.

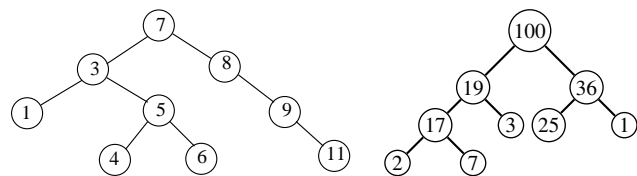
Clearly, this code is *prefix*. Moreover, if we take into account the frequencies of particular symbols of the input alphabet in the input text, we obtain a *lossless data compression*.

often, they are called a *child* and a *parent* (motivated by the genealogical trees).

The most common data structures are the *binary trees*, which are special cases of a rooted trees: there, every vertex has at most two children (sometimes, the term *binary tree* implies that every vertex is either a leaf, or has *exactly* two children; to avoid ambiguity, such trees are often called full binary trees). Such trees are very useful in search procedures. If the vertices are associated to keys from a totally ordered set (eg. the integers), the search for the vertex with a given key is performed by searching the path from the root of the tree to that vertex. At every vertex, compare its key to the desired one. This decides whether one continues to the left or to the right, or stop the search if it is found. If this algorithm is to be correct, one of the children with all its successors must have lower keys than the keys of the other child and all its successors.

In order for the search to be efficient, some effort must be made to keep the binary trees *balanced*, with the lengths of the paths from the root to the leaves differing by at most one. The most unfortunate example of a binary tree on  $n$  vertices is the path graph  $P_n$  (which may be formally considered a binary tree), while the most desired case is the perfect complete binary tree, where every vertex that is not a leaf has exactly two children, and all leaves are at the same level. Such a tree can be constructed only when the number of vertices is of the form  $n = 2^k - 1$ ,  $k = 1, 2, \dots$ . Therefore, in a balanced tree, finding the vertex with a given key value can be done in  $O(\log_2 n)$  steps. Such trees are often called *binary search trees*. Think out as an exercise how to efficiently perform basic graph operations over binary search trees (additions and removals of the vertex with a given key as well as how to keep the tree balanced).

An extraordinarily useful example of binary trees is the structure of a *heap*. It is a full balanced binary tree, where the keys are either strictly decreasing along each path from the root (the so called *max heap*), or they are increasing (the *min heap*). Because of this ordering along the paths in a max heap, the maximum key value of the heap can be found in constant time and removed in logarithmic time (similarly with minimum in the min heap). The desired maximum is just at the root and after removing it we need to balance the shape of the heap. Prove this is possible in logarithmic time yourself!



The left-hand diagram shows a binary search tree. In the right-hand diagram, there is a max heap.

Much literature is devoted to trees, their applications and miscellaneous variations.

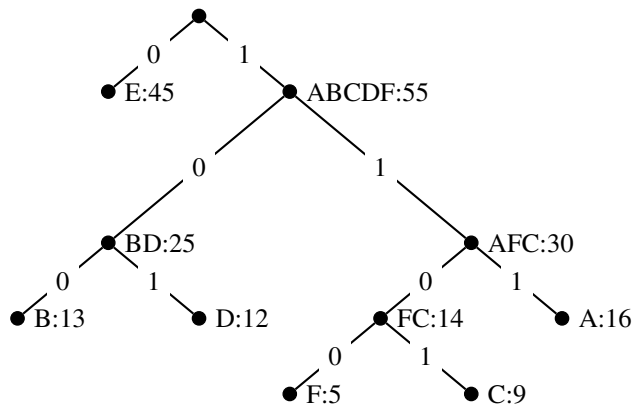
Let  $M$  be the list of frequencies of the symbols of the input alphabet in the input text. The algorithm constructs the optimal binary tree (the so-called *minimum-weight binary tree*) and the assignment of the symbols to the leaves.

- Select the two least frequencies  $w_1, w_2$  from  $M$ . Create a tree with two leaves labeled by the corresponding symbols and root labeled by  $w_1 + w_2$ , then replace the values  $w_1, w_2$  with the new value  $w_1 + w_2$  in  $M$ .
- Repeat the above step; if the selected value from  $M$  is a sum, then simply “connect” the existing subtree.
- The code of each symbol is determined by the path from the root to the corresponding leaf (left edge = “0”, right edge = “1”, for instance).

**13.B.18.** Find the Huffman code for the input alphabet with the frequencies [‘A’:16, ‘B’:13, ‘C’:9, ‘D’:12, ‘E’:45, ‘F’:5].

**Solution.** If we naively assign a 3-bit code to each letter of the alphabet, then this message of length 100 consumes 300 bits.

We show that Huffman code is more succinct. We build the tree according to the algorithm.



We have thus obtained the codes  $A : 111, B : 100, C : 1101, D : 101, E : 0, F : 1100$ . Multiplying the code lengths by the frequencies, we can see that a 100-letter message with the given distribution of letters is encoded into only

$$3 \cdot 16 + 3 \cdot 13 + 4 \cdot 9 + 3 \cdot 12 + 1 \cdot 45 + 4 \cdot 5 = 224 \text{ bits.} \quad \square$$

**C. Minimum spanning tree**

**13.C.1.** How many spanning trees (see 13.2.6) of the graph  $K_5$  are there? And how many are there if we do not distinguish isomorphic ones?

**Solution.** There are three pairwise non-isomorphic spanning trees (with degree sequences  $(1, 2, 2, 2, 1), (1, 2, 3, 1, 1), (4, 1, 1, 1, 1)$ ). The corresponding classes of isomorphic spanning trees have  $5 \cdot \binom{4}{2} \cdot 2, 5 \cdot 4 \cdot 3,$  and  $5$  elements, respectively. Altogether, there are  $125 = 5^3$  spanning trees, which is in accordance with Cayley’s formula for the number of spanning trees of a complete graph (see 13.4.11).  $\square$

**13.1.17. Remarks on sorting.** Suppose it is required to distinguish all different sortings of  $n$  elements, thus distinguishing among  $n!$  different objects. If there is no information other than comparing the order of two single elements, then the tree of all possible decision paths can be written down. The sorting provides a path through this binary tree. As seen, any binary tree of depth  $h$  (i.e.  $h - 1$  is the length of the longest path from the root to a leaf) has at most  $2^{h-1}$  leaves. It follows that a tree of depth  $h$  satisfying  $2^{h-1} \geq n!$  is needed.



Consequently the depth  $h$  satisfies  $h \log 2 > \log n!$ .

$$\begin{aligned} \log n! &= \log 1 + \log 2 + \dots + \log n \\ &> \int_1^n \log x \, dx = n \log n - (n - 1) \\ h &> \frac{n \log n - n}{\log 2} > n \log n - n \end{aligned}$$

It is proved that the depth of the necessary binary tree is bounded from below by an expression of size  $n \log n$ . Hence no algorithm based only on the comparison of two elements of the ordered set can have a better worst case run than  $O(n \log n)$ .

The latter claim is not true if there is further relevant information. For example, if it is known that only a finite number of  $k$  values may appear among our  $n$  elements, then one may simply run through the list counting how many occurrences of the individual values are there, and hence write the right ordered list from scratch. This all happens in linear time!

**13.1.18. Tree isomorphisms.** Simple features of trees are exploited in order to illustrate the (generally difficult) problem of graph isomorphisms on this special class of graphs.



First, strengthen the structure to be preserved by the isomorphisms. Then show that the obtained procedure is also applicable to the most general trees.

In order to keep more information about the structure of rooted trees, remember the relations parent–child. Also have the children of every node sorted in a specific order (for instance, from left to right if drawn on a diagram). Such trees are called *ordered trees* or *plane trees*. They are formally defined as a tuple  $T = (V, E, v_r, \nu)$ , where  $\nu$  is a partial order on the edges such that a pair of edges is comparable if and only if they have the same tail (i.e. they all go from one parent vertex to all its children).

A homomorphism of rooted trees  $T = (V, E, v_r)$  and  $T' = (V', E', v'_r)$  is a graph morphism  $\varphi : T \rightarrow T'$  such that  $v_r$  is mapped to  $v'_r$ ; similarly for isomorphisms. For plane trees, it is further required that the morphism preserves the partial orders  $\nu$  and  $\nu'$ .

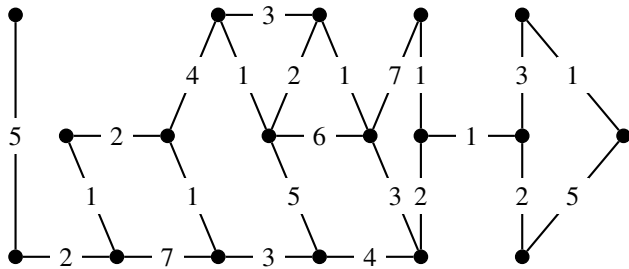
**13.C.2.** Let the vertices of  $K_6$  be labeled  $1, 2, \dots, 6$  and let every edge  $\{i, j\}$  be assigned the integer  $[(i + j) \bmod 3] + 1$ . How many minimum spanning trees are there in this graph?

**Solution.** There are five edges whose value is 1: four of them lie on the cycle 12451 and the remaining one is the edge 36. Therefore, they form a disconnected subgraph of the complete graph, so the spanning tree must contain at least one edge of value 2. Thus, the total weight of a minimum spanning tree is at least  $4 \cdot 1 + 2 = 6$ . And indeed, there exist spanning trees with this weight. We select all the edges of value 1 except for one that lies on the mentioned cycle and connect the resulting components 1245 and 36 with any edge of value 2. There are four such edges. Altogether, there are  $4 \cdot 4 = 16$  minimum spanning trees.  $\square$

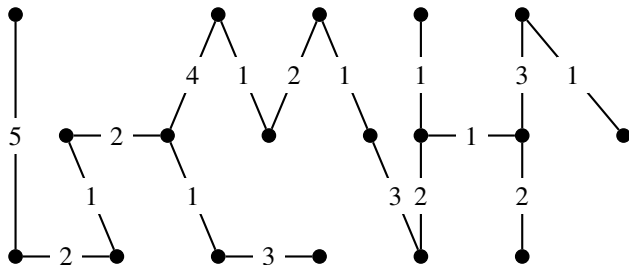
**13.C.3.** Find a minimum spanning tree of the following graph using

- i) Kruskal's algorithm,
- ii) Jarník's (Prim's) algorithm.

Explain why we cannot apply Borůvka's algorithm directly.



**Solution.** The spanning tree is



Borůvka's algorithm cannot be applied directly since the mapping of the weights to the edges is not injective. However, this can be fixed easily by slight modifications of the weights.  $\square$

**13.C.4.** Consider the following procedure for finding the shortest path between a given pair of vertices in an undirected weighted graph: First, we find a minimum spanning tree. Then, we proclaim that the only path between the pair of vertices in the obtained spanning tree is the shortest one. Prove the correctness of this method, or disprove it by providing a counterexample.  $\circ$

**13.C.5.** We are given the following table of distances of world metropolises: London, Mexico City, New York, Paris,  $\square$

CODING THE PLANE TREES

Given the plane tree  $T = (V, E, v_r, \nu)$ . It has a code  $W$  by strings of ones and zeros, defined recursively as follows:

Start with the word 01 for the root  $v_0$  and write  $W = 0W_1 \dots W_\ell 1$ , where  $W_i$  are the  $\ell$  still unknown words for the subtrees rooted by the children of  $v_0$ . In particular the code of the tree with just one vertex is  $W = 01$ .

Applying the same procedure recursively over the children and concatenating the results defines the code.

The tree in the left-hand diagram above (in the end of 13.1.16) is encoded as follows (the children of a vertex are ordered from left to right,  $W_r$  is for the code of the child with key  $r$ ):

$$\begin{aligned} 0W_3W_81 &\mapsto 00W_1W_510W_911 \\ &\mapsto 00010W_4W_61100W_{11}111 \\ &\mapsto 000100101110001111. \end{aligned}$$

Imagine drawing the entire plane tree with one move of the pencil, starting with an arrow ending in the root and going downwards with arrows towards the leaves and then upwards to the predecessors, reaching consecutively all the leafs from the left to the right and writing 0 when going down and 1 when going up. The very last arrow is then leaving the root upwards.

**Theorem.** Two plane trees are isomorphic if and only if their codes are the same.

**PROOF.** By construction, two isomorphic trees are assigned the same code. It remains to show that different codes lead to non-isomorphic plane trees.

This is proved by induction on the length of the code (i.e. the number of zeros and ones). This length is  $2(|E| + 1)$ , (twice the number of vertices; therefore, the proof can be viewed as an induction on the number of vertices of the tree  $T$ ). The shortest code corresponds to the smallest tree on one vertex. Assume that the proposition holds for all trees up to  $n$  vertices, i.e. for codes of length up to  $k = 2n$ , and consider a code of the form  $0W1$ , where  $W$  is a code of length  $2n$ . Find the shortest prefix of  $W_1$  which contains the same number of zeros and ones (when drawing a diagram of the tree, this is the first moment when we return to the root of the tree that corresponds to the code  $0W1$ ). Similarly, find the next part of the code  $W$  that contains the same number of zeros and ones, etc. Hence the code  $W$  can be written as  $W = W_1W_2 \dots W_\ell$ . By the induction hypothesis, the codes  $W_i$  correspond uniquely (up to isomorphism) to plane trees, and the order of their roots, being the children of our tree  $T$ , is given uniquely by the order in the code. Therefore, the tree  $T$  is determined uniquely by the code  $0W1$  up to isomorphism.  $\square$

Peking, and Tokyo:

$$\begin{pmatrix} & L & MC & NY & P & Pe & T \\ L & & 5558 & 3469 & 214 & 5074 & 5959 \\ MC & & & 2090 & 5725 & 7753 & 7035 \\ NY & & & & 3636 & 6844 & 6757 \\ P & & & & & 5120 & 6053 \\ Pe & & & & & & 1307 \end{pmatrix}$$

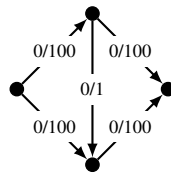
What is the least total length of wire used for interconnecting these cities (assuming the length necessary to connect a given pair of cities is equal to the distance in the table). ○

13.C.6. Using the matrix variation of the JarnÅk-Prim algorithm, find a minimum spanning tree of the graph given by the following matrix:

$$\begin{pmatrix} - & 12 & - & 16 & - & - & - & 13 \\ 12 & - & 16 & - & - & - & 14 & - \\ - & 16 & - & 12 & - & 14 & - & - \\ 16 & - & 12 & - & 13 & - & - & - \\ - & - & - & 13 & - & 14 & - & 15 \\ - & - & 14 & - & 14 & - & 15 & - \\ - & 14 & - & - & - & 15 & - & 14 \\ 13 & - & - & - & 15 & - & 14 & - \end{pmatrix}$$

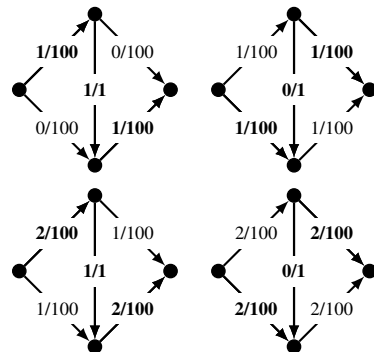
D. Flow networks

13.D.1. An example of bad behavior of the Ford-Fulkerson algorithm. The worst-case time complexity of the Ford-Fulkerson algorithm is  $O(E \cdot |f|)$ , where  $|f|$  is the size of a maximum flow. Consider the following network:



The bad behavior of the algorithm is due to the fact that it uses depth-first search to find unsaturated paths.

Solution. We proceed strictly by depth-first search (examining the vertices from left to right and then top down):



Use the encoding of plane trees to encode any tree. Deal first with the case of rooted trees. Determine the order of the children of every vertex uniquely up to isomorphism. The order is unimportant if and only if the subgraphs determined by the respective children are isomorphic.



The same construction can be used as for the plane trees, ordering the vertices lexicographically with respect to their codes (see 12.6.5 for the related concepts). This means that codes  $W_1, W_2$  satisfy  $W_1 > W_2$  if and only if  $W_1$  contains a one at an earlier position than  $W_2$  or  $W_2$  is a prefix of  $W_1$ . The rooted tree as a whole is described by the same recursive procedure: if the children of a vertex  $v$  are coded by  $W_1, \dots, W_\ell$ , then the code of the vertex  $v$  is

$$0W_1 \dots W_\ell 1,$$

where the order is selected so that  $W_1 \leq W_2 \leq \dots \leq W_\ell$ .

If no vertex is designated to be the root of a tree, the root can be designed so that it would be almost “in the middle” of the tree. This can be realized by assigning an integer to every vertex of the tree which describes its *eccentricity*. That eccentricity  $ex_T(v)$  of a vertex  $v$  in a graph  $T$  is defined to be the greatest possible distance between  $v$  and some vertex  $w$  in  $T$ . This concept is meaningful for all graphs; however, by the absence of cycles in trees, it is guaranteed that there are at most two vertices with the minimal eccentricity. ○

Lemma. Let  $C(T)$  be the set of those vertices of a tree  $T$  whose eccentricity is minimal. Then,  $C(T)$  contains either a single vertex, or exactly two vertices, which are connected with an edge in  $T$ .

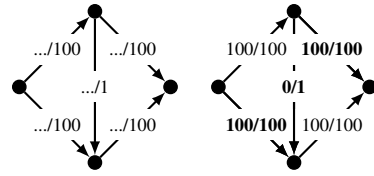
PROOF. The claim is proved by induction, using the trivial fact that the most distant vertex from any vertex  $v$  must be a leaf. Therefore, the center of  $T$  coincides with the center of the tree  $T'$  which is created from the tree  $T$  by removing all its leaves and the corresponding edges. After a finite number of such steps, there remains either just one vertex, or a subtree with two vertices. □

$C(T)$  determined by the latter lemma is called the *center of the graph*, and the minimal eccentricity is called the *radius of the graph*.

A unique (up to isomorphism) code can now be assigned to every tree. If the center of  $T$  contains only one vertex, use it as the root. Otherwise, create the codes for the two rooted subtrees of the original tree without the edge that connects the vertices of the center, and the code of  $T$  is the code of the rooted tree  $(T, x)$ , where  $x$  is the vertex of the center whose subtree has lexicographically smaller code.

Corollary. Trees  $T$  and  $T'$  are isomorphic if and only if they are assigned the same code.

The above ideas imply that the algorithm for verifying planar tree isomorphism can be implemented in linear time with respect to the number of vertices of the trees.



We can see that 200 iterations were needed in order to find the maximum flow.  $\square$

**13.D.2.** Find the size of a maximum flow in the network given by the following matrix  $A$ , where vertex 1 is the source and vertex 8 is the sink. Further, find the corresponding minimum cut.

$$A = \begin{pmatrix} - & 16 & 24 & 12 & - & - & - & - \\ - & - & - & - & 30 & - & - & - \\ - & - & - & - & 9 & 6 & 12 & - \\ - & - & - & - & - & - & 21 & - \\ - & - & - & - & - & 9 & - & 15 \\ - & - & - & - & - & - & - & 9 \\ - & - & - & - & - & - & - & 18 \\ - & - & - & - & - & - & - & - \end{pmatrix}$$

**Solution.** The following augmenting semipaths are found:

- 1–2–5–8 with residual capacity 15.
- 1–2–5–6–8 with residual capacity 1.
- 1–3–5–6–8 with residual capacity 8.
- 1–4–7–8 with residual capacity 12.
- 1–3–7–8 with residual capacity 6.

The total size of the found flow is 42. We can see that it is indeed of maximum size from the fact the cut consisting of edges  $(5, 8)$ ,  $(6, 8)$ , and  $(7, 8)$  has also size 42 (and it is thus a minimum cut).  $\square$

**13.D.3.** The following picture depicts a flow network (the numbers  $f/c$  define the actual flow and the capacity of a given edge, respectively). Decide whether the given flow is maximum. If so, justify your answer. If not, find a maximum flow and describe the used procedure in detail. Find a minimum cut in the network.

picture skipped

**Solution.** In the given network, There exists an augmenting (semi)path 1–2–3–4–8 with residual capacity 4. Its saturation results in a flow of size 32. Since the cut  $(3, 8)$ ,  $(5, 8)$ ,  $(2, 4)$ ,  $(6, 4)$  is of the same size, we have found a maximum flow.  $\square$

**13.D.4.** Find a maximum flow and a minimum cut in the following flow network (source=1,sink=14).

The trees form a special class of graphs. They are often used in miscellaneous variations and with additional requirements. We return to them later, in connection with practical applications.

Now follows another extraordinarily important class of graphs.

**13.1.19. Planar graphs.** Some graphs are drawn in the

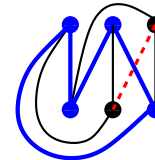


plane in such a way that their edges do not “cross” one another. This means that every vertex of the graph is identified with a point of the plane, and an edge between vertices  $v$ ,  $w$  corresponds to a continuous curve  $c : [0, 1] \rightarrow \mathbb{R}^2$  that connects the vertices  $c(0) = v$  and  $c(1) = w$ . Furthermore, suppose that edges may intersect only at their endpoints. This describes a *planar graph*  $G$ .

The question whether a given graph admits a realization as a planar graph often emerges in practical problems. Here is an example:

Providers of water, electricity, and gas have their connection spots near three houses (each provider has one spot). Each house wants to be connected to each resource so that the connections would not cross (they might want not to dig too deep, for instance). Is it possible to do this? The answer is: “no”.

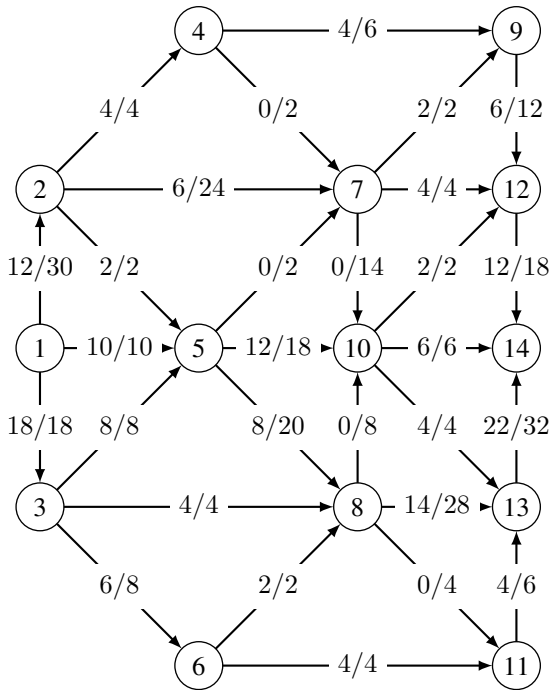
In this particular case, it is clear from the diagram. There is a complete bipartite graph  $K_{3,3}$ , where three of the vertices correspond to the connection spots, while the other three represent the houses. The edges are the connections between the spots and the houses. All edges can be placed except the last one – see the diagram, where the dashed edge cannot be drawn without crossing any more:



For a complete proof, more mathematical tools are needed. A complete explanation is not provided here, but an indication of the reasoning follows.



One of the basic results from the topology of the plane (the *Jordan curve theorem*) states that every closed continuous curve  $c$  in the plane that is not self-intersecting (i.e. it is a “crooked circle”) divides the plane into two distinct parts. In other words, every other continuous curve which connects a point inside a curve  $c$  and a point outside  $c$  must intersect  $c$ . If the edges are realized as piecewise linear curves (every edge composed of finitely many adjacent line segments), then it is quite easy to prove the Jordan curve theorem (you might do it yourself!). The general theorem can be proved by approximating the continuous curves by piecewise linear ones (quite difficult to do, but it is much easier if the curve is assumed to be piecewise differentiable).



**Solution.** The paths are saturated in the following order:

$$\begin{aligned}
 &1 \xrightarrow{18} 2 \xrightarrow{18} 7 \xrightarrow{14} 10 \xleftarrow{12} 5 \xrightarrow{12} 8 \xrightarrow{4} 11 \xrightarrow{2} 13 \xrightarrow{10} 14 \\
 &1 \xrightarrow{16} 2 \xrightarrow{16} 7 \xrightarrow{12} 10 \xleftarrow{10} 5 \xrightarrow{10} 8 \xrightarrow{14} 13 \xrightarrow{8} 14
 \end{aligned}$$

We have found a flow of size 50. And indeed, it is a maximum flow since there is no further unsaturated path. If we look for the reachable vertices, we can also find a cut with capacity 50, consisting of edges

$$[2, 4] : 4, [7, 9] : 2, [7, 12] : 4, [10, 12] : 2, [10, 14] : 6, [13, 14] : 32. \quad \square$$

**13.D.5.** Find a maximum flow in the following network on the vertex set  $\{1, 2, \dots, 9\}$  with source 1 and sink 9 using the Ford-Fulkerson algorithm (during the depth-first search, choose the vertices in ascending order). Find a minimum cut in this network. Describe the steps of the procedure in detail. The edges  $e \in E$  as well as the lower and upper bounds on the flow ( $l(e)$  and  $u(e)$ ) and the current flow  $f(e)$  are given in the table:

Consider the graph  $K_{3,3}$ . The triples of vertices that are not connected with edges are indistinguishable up to order. Therefore the thick cycle can be considered the general case of a cycle with four points in the graph. The position of the remaining two vertices can then be discussed. In order for the graph to be planar, either both of the vertices must lie inside the cycle, or both outside. Again, these possibilities are equivalent, so it can be assumed without loss of generality that they are in opposite sides, as are the black vertices on the diagram. Now, their position with respect to a suitable cycle with two thick edges and two thin black edges can be discussed (i.e. through three gray vertices and one black one). Then, we can discuss the position of the remaining black vertex with respect to this cycle. This leads to the impossibility of drawing the last (dashed) edge without crossing the thick cycle.

It can be shown similarly that the complete graph  $K_5$  is not planar either. We provide a pure combinatorial argument why  $K_5$  and  $K_{3,3}$  cannot be planar graphs below, see the Corollary in the end of the next subsection.

Notice that if a graph  $G$  is “expanded” by dividing some of its edges (i.e. adding new vertices in the edges), then the new graph is planar if and only if  $G$  is planar. The new graph is a *subdivision* of  $G$ . Planar graphs must not contain any subdivision of  $K_{3,3}$  or  $K_5$ . The reverse implication is also true:

KURATOWSKI THEOREM

**Theorem.** A graph  $G$  is planar if and only if none of its subgraphs is isomorphic to a subdivision of  $K_{3,3}$  or  $K_5$ .

The proof is complicated, so it is not discussed here. Much attention is devoted to planar graphs both in research and practical applications.

There are algorithms which are capable of deciding whether or not a given graph is planar in linear time. Direct application of the Kuratowski theorem would lead to a worse time complexity.

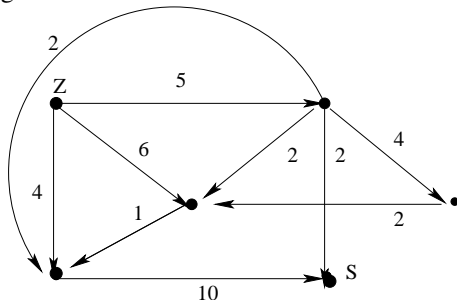
**13.1.20. Faces of planar graphs.** Consider a planar graph  $G$  embedded in the plane  $\mathbb{R}^2$ . Let  $S$  be the set of those points  $x \in \mathbb{R}^2$  which do not belong to any edge of the graph (nor are vertices). In this way, the set  $\mathbb{R}^2 \setminus G$  is partitioned into connected subsets  $S_i$ , called the *faces of the planar graph*  $G$ . Since the graphs are finite, there is exactly one unbounded face  $S_0$ . The set of all faces are denoted by  $S = \{S_0, S_1, \dots, S_k\}$ , and the planar graph by  $G = (V, E, S)$ .

The simplest case of a planar graph is a tree. Every tree is a planar graph since it can be constructed by step-by-step addition of leaves, starting with single vertex. Of course, the Kuratowski theorem can also be applied—when there is no cycle in a graph  $G$ , then there cannot be a subdivision of  $K_{3,3}$  or  $K_5$ , either. Since a tree  $G$  cannot contain a cycle, there is

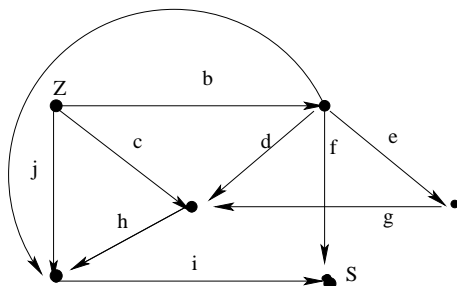


$e$	$l(e)$	$u(e)$	$f(e)$
(1,2)	0	6	0
(1,3)	0	6	0
(1,6)	0	4	0
(2,3)	0	2	0
(2,4)	0	3	0
(3,4)	0	4	0
(3,5)	0	4	0
(4,5)	3	5	4
(4,8)	0	3	0
$e$	$l(e)$	$u(e)$	$f(e)$
(5,1)	0	3	0
(5,6)	0	6	0
(5,7)	0	5	4
(5,8)	0	5	0
(6,9)	0	5	0
(7,4)	1	6	4
(7,9)	0	3	0
(8,9)	0	9	0

**13.D.6.** A cut in a network  $(V, E, s, t, w)$  can also be viewed as a set  $C \subset E$  of edges such that in the network  $(V, E \setminus C, s, t, w)$ , there is no path from the source  $s$  to the sink  $t$ , but if any edge  $e$  is removed from  $C$ , then the resulting set does not satisfy this property, i. e., there is a path from  $s$  to  $t$  in  $(V, E \setminus C \cup e, s, t, w)$ . Find all cuts (and their sizes) in the following network:



**Solution.** Let us fix the following edge labeling:



Then, there are cuts:  $\{f,i\}, \{f,h,j,a\}, \{f,j,c,a,d,e\}, \{f,j,c,a,d,g\}, \{b,j,c\}, \{b,j,h\}, \{b,i\}$ . Their capacities are 12, 9, 20, 18, 15, 10, and 15, respectively.  $\square$

**13.D.7.** Find a maximum flow in the network given in the above exercise.  $\circ$

only one face  $S_0$  there (the unbounded face). Since the number of edges of a tree is related to the number of its vertices, cf. the formula 13.1.15(5), it follows that

$$|V| - |E| + |S| = 2$$

for all trees.

Surprisingly, the latter formula linking the number of edges, faces, and vertices can be derived for all planar graphs. The formula is named after Leonhard Euler. Especially, the number of faces is independent of the particular embedding of the graph in the plane:

EULER'S FORMULA

**Theorem.** Let  $G = (V, E, S)$  be a connected planar graph. Then,

$$|V| - |E| + |S| = 2.$$

**PROOF.** The proof is by induction on the number of edges. The graph with zero or one edge satisfies the formula. Consider a graph  $G$  with  $|E| > 1$ . If  $G$  does not contain a cycle, then it is a tree, and the formula is already proved for this case.

Suppose that there is an edge  $e$  of  $G$  that is contained in a cycle. Then, the graph  $G' = G \setminus e$  is connected, and it follows from the induction hypothesis that  $G'$  satisfies Euler's formula:

$$|V| - (|E| - 1) + (|S| - 1) = 2,$$

since removing an edge necessarily leads to merging two faces of  $G$  into one face in  $G'$ . Hence Euler's formula is valid for the graph  $G$ .  $\square$

**Corollary.** Let  $G = (V, E, S)$ , be a planar graph with  $|V| = n \geq 3$ , and  $|E| = e$ . Then

- There is the inequality  $e \leq 3n - 6$  which becomes equality if and only if  $G$  is a maximal planar graph (adding any edge to  $G$ , would violate planarity).
- If  $G$  does not contain a triangle (i.e. the graph  $K_3$  is not a subgraph), then  $e \leq 2n - 4$ .

**PROOF.** Continue adding edges to a given graph until it is maximal. If the obtained maximal graph  $G$  satisfies the inequality with equality, then the inequality holds for the original graph as well.

Similarly, if the graph  $G$  is not connected, two of its components can be connected with a new edge, so such a graph cannot be maximal. Even if it were connected but not 2-connected, there would exist a vertex  $v \in V$  such that when it is removed, the graph  $G$  collapses into several components  $G_1, \dots, G_k, k \geq 2$ . However, then an edge can be added between these components without destroying the planarity of the original graph  $G$  (draw a diagram!). Therefore, it can be assumed from the beginning that the original graph  $G$  is a maximal planar 2-connected graph.

As shown in theorem 13.1.11, every 2-connected graph can be constructed from the triangle  $K_3$  by splitting edges and attaching new ones. It is easily proved by induction that every



Further exercises on maximum flows and minimum cuts can be found on page 1220.

**E. Classical probability and combinatorics**

In this section, we recall the methods we learned as early as in the first chapter.

**13.E.1.** We throw  $n$  dice. What is the probability that none of the values 1, 3, 6 is cast?

**Solution.** We can also see the problem as throwing one dice  $n$  times. The probability that none of the values 1, 3, 6 is cast in the first throw is  $1/2$ . The probability that they are cast neither in the first throw nor in the second one is clearly  $1/4$  (the result of the first throw has no impact on the result of the second one). Since this holds generally, i. e., the results of different throws are (stochastically) independent, the wanted probability is  $1/2^n$ .  $\square$

**13.E.2.** We have a pack of ten playing cards, exactly one of which is an Ace. Each time, we randomly draw one of the ten cards and then put it back. How many times do we have to repeat this experiment if we require the probability of getting the Ace at least once to be at least 0.9?

**Solution.** Let  $A_i$  be the event “the Ace is picked in the  $i$ -th draw”. The events  $A_i$  are (stochastically) independent. Hence, we know that

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdot \dots \cdot (1 - P(A_n))$$
 for every  $n \in \mathbb{N}$ . We are looking for an  $n \in \mathbb{N}$  such that

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - (1 - P(A_1)) \cdot (1 - P(A_2)) \cdot \dots \cdot (1 - P(A_n)) > 0.9.$$

Apparently, we have  $P(A_i) = 1/10$  for any  $i \in \mathbb{N}$ . Therefore, it suffices to solve the inequality

$$1 - \left(\frac{9}{10}\right)^n > 0.9,$$

whence

$$n > \frac{\log_a 0.1}{\log_a 0.9}, \quad \text{where } a > 1.$$

Evaluating this, we find out that we have to repeat the experiment at least 22 times.  $\square$

**13.E.3.** We randomly draw six cards from a pack of 32 cards (containing four Kings). Calculate the probability that the sixth card is a King and, at the same time, it is the only King drawn.

**Solution.** By the theorem on product of probabilities, the result is

$$\frac{28}{32} \cdot \frac{27}{31} \cdot \frac{26}{30} \cdot \frac{25}{29} \cdot \frac{24}{28} \cdot \frac{4}{27} \doteq 0.0723. \quad \square$$

**13.E.4.** We randomly draw two cards from a pack of 32 cards (containing four Aces). Calculate the probability that the second card drawn is an Ace if:

- a) the first card is put back; b) the first card is not put back.

**Solution.** If the first card is put back in the pack, then we clearly repeat an experiment with 32 possible outcomes (with the same probability), 4 of which are favorable. Therefore,

face of a planar graph must be bounded by a cycle (which seems intuitively apparent).

However, if there is a face of our maximal planar graph  $G$  that is not bounded by a triangle, then this face can be split with another edge (a “diagonal” in geometrical terminology), so  $G$  would not be maximal. It follows that all faces of  $G$  are bounded by triangles  $K_3$ . Hence  $3|S| = 2|E|$ .

It suffices to substitute  $|S| = \frac{2}{3}|E|$  for the number of faces in Euler’s formula.

The second proposition is analogous; now, the faces of the maximal planar graph without triangles are bounded by either four or five edges, whence it follows that  $4|S| \leq 2|E|$  with the equality if and only if there are just quadrangles there.  $\square$

The corollary implies (even without the Kuratowski theorem) that neither  $K_5$  nor  $K_{3,3}$  is planar: in the former case,  $|V| = 5$  and  $|E| = 10 > 3|V| - 6$ , while in the latter,  $|V| = 6, |E| = 9 > 2|V| - 4$ , which is again a contradiction since  $K_{3,3}$  does not contain a triangle.

**13.1.21. Convex polyhedra in the space.** Planar graphs can



be imagined as drawn on the sphere instead in the plane. The sphere can be constructed from the plane by attaching one point “at infinity”. Again, faces of such graphs can be discussed, and the faces are now equivalent to one another (even the face  $S_0$  is bounded).

On the contrary, every convex polyhedron  $P \subseteq \mathbb{R}^3$  can be imagined as a graph drawn on the sphere (project the vertices and edges of the polyhedron onto a sufficiently large sphere from any point inside  $P$ ). Dropping a point inside one of the faces (that face becomes the unbounded face  $S_0$ ) then leads to the planar graph as above – the sphere with the hole is spread in the plane.

The planar graphs that are formed of convex polyhedra are clearly 2-connected since every pair of vertices of a convex polyhedron lies on a common cycle. Moreover, every face is interior to its boundary cycle and the graphs of convex polyhedra are always 3-connected.

In fact, they are just such graphs as the following Steinitz’s theorem says (we omit the proof):

STEINITZ’S POLYHEDRA THEOREM

**Theorem.** A graph  $G$  is the graph of a convex polyhedron if and only if it is planar and 3-vertex-connected.



**13.1.22. Platonic solids.** As an illustration of the combinatorial approach to polyhedral graphs, we classify all the regular polyhedra. These are those built up from one type of regular polygons so that the same number of them touch at every vertex. It was known as early as in the epoch of the ancient philosopher Plato that there are only five of them:

the wanted probability is  $1/8$ . However, even if we do not put the first card back, the probability is the same. Clearly, the probability of a given card being drawn the first time is the same as for the second time. Of course, we can also apply the conditional probability. This leads to

$$\frac{4}{32} \cdot \frac{3}{31} + \frac{28}{32} \cdot \frac{4}{31} = \frac{1}{8}. \quad \square$$

**13.E.5. Combinatorial identities.** Use combinatorial means to derive the following important identities (in particular, do not use induction):

Arithmetic series  $\sum_{k=0}^n k = \frac{n(n+1)}{2} = \binom{n+1}{2}$

Geometric series  $\sum_{k=0}^n x^k = \frac{x^{n+1}-1}{x-1}$

Binomial theorem  $(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$

Upper binomial theorem  $\sum_{k=0}^n \binom{k}{m} = \binom{n+1}{m+1}$

Vandermonde's convolution<sup>1</sup>  $\binom{m+n}{r} = \sum_{k=0}^r \binom{m}{k} \binom{n}{r-k}$ .

○

**13.E.6.** An urn consists of 30 red balls and 70 green balls. What is the probability of getting exactly  $k$  red balls in a sample of size 20,  $0 \leq k \leq 20$  if the sampling is done with replacement (repetition allowed)?

**Solution.** Any time we take a sample from the urn we put it back before the next sample (sampling with replacement). Thus in this experiment each time we sample, the probability of choosing a red ball is  $\frac{30}{100}$  and we repeat this in 20 independent trials. Thus, using the binomial formula, we obtain

$$P(k \text{ red balls}) = \binom{20}{k} (0.3)^k (0.7)^{20-k}$$

□

**13.E.7.** An urn consists of 30 red balls and 70 green balls. What is the probability of getting exactly  $k$  red balls in a sample of size 20,  $0 \leq k \leq 20$  if the sampling is done without replacement (repetition not allowed)?

**Solution.** Let  $A$  be the event of getting exactly  $k$  red balls. To find  $P(A)$  we need to find  $|A|$  and the total number of possibilities  $|S|$ . Here  $|S| = \binom{100}{20}$ . Next, to find  $|A|$ , we need to find out in how many ways one can choose  $k$  red balls and  $20 - k$  green balls. Thus,

$$|A| = \binom{30}{k} \binom{70}{20-k}.$$

Thus,

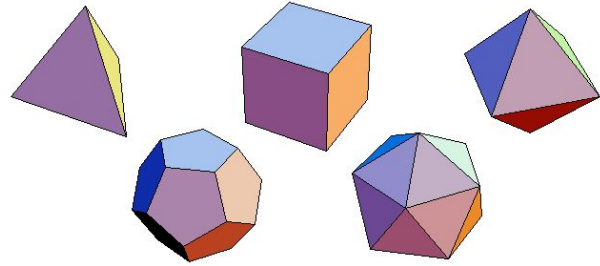
$$P(A) = \frac{|A|}{|S|} = \frac{\binom{30}{k} \binom{70}{20-k}}{\binom{100}{20}}.$$

□

**13.E.8.** Assume that there are  $k$  people in a room and we know that

- $k = 5$  with probability  $\frac{1}{4}$ ;
- $k = 10$  with probability  $\frac{1}{4}$ ;
- $k = 15$  with probability  $\frac{1}{2}$ .

<sup>1</sup>Also called hockey identity.



Translate the condition of regularity to the properties of the corresponding graphs: Every vertex needs the same degree  $d \geq 3$ , and the boundary of every face must contain the same number  $k \geq 3$  of vertices. Let  $n, e, s$  denote the total number of vertices, edges, and faces, respectively.

Firstly, the relation of the vertex degrees and the number of edges requires

$$dn = 2e.$$

Secondly, every edge lies in the boundary of exactly two faces, so

$$2e = ks.$$

Thirdly, Euler's formula states that

$$2 = n - e + s = \frac{2e}{d} - e + \frac{2e}{k}.$$

Put this together. The constants  $d$  and  $k$  must satisfy

$$\frac{1}{d} - \frac{1}{2} + \frac{1}{k} = \frac{1}{e}.$$

Since  $d, k, e, n$  are positive integers (in particular,  $\frac{1}{e} > 0$ ), this equality restricts the possibilities. Especially, the left-hand side is maximal for  $d = 3$ . Substitute this value, to obtain the inequality

$$-\frac{1}{6} + \frac{1}{k} = \frac{1}{e} > 0.$$

It follows that  $k \in \{3, 4, 5\}$  for a general  $d$ . The roles of  $k$  and  $d$  are symmetric in the original equality, so also  $d \in \{3, 4, 5\}$ . Checking each of the remaining possibilities, yields all the solutions:

$d$	$k$	$n$	$e$	$s$
3	3	4	6	4
3	4	8	12	6
4	3	6	12	8
3	5	20	30	12
5	3	12	30	20

It remains to show that the corresponding regular polyhedra exist. This is already seen in the above diagrams, but that is not a mathematical proof. The existence of the first three is apparent. Concentrate on the geometrical construction of the regular dodecahedron (draw a diagram!).

Begin with a cube, building "A-tents" on all its sides simultaneously. The upper horizontal poles are set on the level of the cube's sides so that those of adjacent sides are perpendicular to each other. Its length is chosen so that the trapezoids of the lateral sides would have three sides of the same length. Now, simultaneously



- i) What is the probability that at least two of them have been born in the same month? Assume that all months are equally likely.
- ii) Given that we already know there are at least two people that celebrate their birthday in the same month, what is the probability that  $k = 10$ ?

**Solution.** Let  $A_k$  be the event that at least 2 people out of  $k$  people have birthdays on the same month. We have The complimentary event  $B_k$  is that none out of  $k$  people have birthday on the same month, i.e. there are exactly  $k$  distinct months when these people were born. Thus

$$P(A_k) = 1 - \frac{P_1^k 2}{7} 12^k, \text{ for } 2 \leq k \leq 12.$$

For  $k > 12$ , obviously,  $P(A_k) = 1$ . Therefore the required

$$\begin{aligned} P &= \frac{1}{4}P(A_5) + \frac{1}{4}P(A_{10}) + \frac{1}{2}P(A_{15}) \\ &= \frac{1}{4}\left(1 - \frac{P_1^5 2}{7} 12^5\right) + \frac{1}{4}\left(1 - \frac{P_1^{10} 2}{7} 12^{10}\right) + \frac{1}{2}. \end{aligned}$$

The second part of the question asks for conditional probability  $P(k = 10|A)$ . According to Bayes' rule:

$$\begin{aligned} P(k = 10|A) &= \frac{P(A|k = 10)P(k = 10)}{P(A)} = \frac{P(A_{10})}{4P(A)} \\ &= \frac{1 - \frac{P_1^{10} 2}{7} 12^{10}}{\left(1 - \frac{P_1^5 2}{7} 12^5\right) + \left(1 - \frac{P_1^{10} 2}{7} 12^{10}\right) + 2} \end{aligned}$$

□

**13.E.9. Hat matching problem**  $N$  guests arrive at a party. Each person is wearing a hat. All hats are collected and then randomly redistributed at the departure. What is the probability that at least one person receives his/her own hat?

**Solution.** Let  $A_i$  be the the event that person  $i$  receives own hat. Then the task is to find  $P(E)$ , where  $E = \cup_{i=1}^N A_i$ . To find  $P(E)$  we use inclusion-exclusion principle:

$$\begin{aligned} P(E) &= P(\cup_{i=1}^N A_i) = \sum_{i=1}^N P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \\ &\sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{N-1} P(\cap_{i=1}^N A_i) \end{aligned}$$

Note, that due to symmetry, the probability of any such intersection depends only on the number of sets  $A_i$  in the intersection, so

$$\begin{aligned} \sum_{i=1}^N P(A_i) &= NP(A_1); \\ \sum_{i < j} P(A_i \cap A_j) &= \binom{N}{2} P(A_1 \cap A_2); \\ \sum_{i < j < k} P(A_i \cap A_j \cap A_k) &= \binom{N}{3} P(A_1 \cap A_2 \cap A_3); \\ &\dots \end{aligned}$$

raise all tents while keeping the ratio of the three sides of the trapezoids. There is a position at which the adjacent trapezoid and triangle sides are coplanar. At that position, the regular dodecahedron is created.

Now, the regular icosahedron can be constructed via the so called *dual graph* construction. The dual graph  $G'$  to a planar graph  $G = (V, E, S)$  has vertices defined as the faces in  $S$  and there is an edge between faces  $S_1$  and  $S_2$  if and only if they share an edge (i.e. were neighbours) in  $G$ . Clearly the dual graph to the dodecahedron is the isosahedron. Exactly as the cube and the octohedron are dual, while tetrahedron is dual to itself.

## 2. A few graph algorithms

In this part, we consider several applications of graph concepts and the algorithms built upon them.

**13.2.1. Graph representations.** As already indicated, algorithms are often formulated with the help of the language of graphs.



The concept of an algorithm can be formalized as a procedure dealing with a (directed) graph whose vertices and/or edges are equipped with further information. The procedure consists in walking through the graph along its edges, while processing the information associated to the visited vertices and edges. Of course, processing the information includes also the decision which outgoing edges must be investigated in a further walk, and in which order.

In the case of an undirected graphs, each (undirected) edge can be replaced with a pair of directed edges.

The graph may also be changed during the run of the algorithm, i.e. vertices and/or edges may be added or removed.

In order to execute such algorithms efficiently (usually on a computer), it is necessary to represent the graph in a suitable way. The adjacency matrix representation is one possibility, cf. 13.1.8. There are many other options based on various lists with suitable pointers.

The *edge list* (also the *adjacency list*) of the graph  $G = (V, E)$  consists of two lists  $V$  and  $E$  that are interconnected by pointers so that every vertex points to the edges it is incident to. and every edge points to its endpoints.

The necessary memory to represent the graph as an edge list is  $O(|V| + |E|)$  since every edge is pointed at twice and every vertex is pointed at  $d$  times, where  $d$  is its degree, and the sum of the degrees of all vertices equals twice the number of edges. Therefore, up to a constant multiple, this is an optimal way of graph representation in memory. It is of interest in how the basic graph operations are processed in both representations. By the basic operations, is meant:

- *removal of an edge,*
- *addition of an edge,*
- *removal of a vertex,*
- *addition of a vertex,*
- *splitting an edge with a new vertex.*

Therefore,

$$P(E) = NP(A_1) - \binom{N}{2} P(A_1 \cap A_2) + \binom{N}{3} P(A_1 \cap A_2 \cap A_3) - \dots$$

We need to find  $P(A_1)$ ,  $P(A_1 \cap A_2)$ ,  $P(A_1 \cap A_2 \cap A_3)$ , ... and so on for all numbers of intersecting sets. Obviously,

$$P(A_1) = \frac{|A_1|}{|S|} = \frac{(N-1)!}{N!} = \frac{1}{N},$$

since  $|S| = N!$  and  $|A_1| = (N-1)!$ . Similarly

$$P(A_1 \cap A_2) = \frac{|A_1 \cap A_2|}{N!} = \frac{(N-2)!}{N!} = \frac{1}{N(N-1)};$$

$$P(A_1 \cap A_2 \cap A_3) = \frac{(N-3)!}{N!} = \frac{1}{N(N-1)(N-2)};$$

Thus,

$$P(E) = N \frac{1}{N} - \binom{N}{2} \frac{1}{N(N-1)} + \binom{N}{3} \frac{1}{N(N-1)(N-2)} - \dots + (-1)^{N-1} \frac{1}{N!},$$

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{N-1} \frac{1}{N!},$$

which approaches  $1 - \frac{1}{e}$  as  $N \rightarrow \infty$ . □

**13.E.10. Texas Hold 'em Poker.** Now, we solve several simple problems about one of the most popular card games—Texas Hold'em Poker. We do not present its rules; they can be easily found on the Internet. What is the probability that:

- i) we are dealt a pair?
- ii) we are dealt an Ace?
- iii) we have one of the six best poker combinations at the end?
- iv) we win if we are holding an Ace and a Three and there are three Twos and the differently-suited Ace on the table (the river has not been dealt yet)?

**Solution.**

- i) There are 4 cards of each of 13 ranks. Therefore, there are  $13 \binom{4}{2} = 78$  pairs. The total number of pairs is  $\binom{13 \cdot 4}{2} = 1326$ . Thus, the wanted probability is  $\frac{1}{17} \doteq 0.06$ .
- ii) One of the cards is an Ace (there are four possibilities) and the other one is arbitrary (51 possibilities). However, this includes the  $\binom{4}{2} = 6$  pairs of Aces twice. Therefore, the number of favorable cases is only  $4 \cdot 51 - 6 = 198$  and the wanted probability is  $\frac{198}{1326} \doteq 0.15$ .
- iii) We compute the probabilities of the particular combinations when dealt five cards at random:  
**ROYAL FLUSH:** There is exactly one such combination for each suit—four in total. Further, there are  $\binom{52}{5} = 2598960$  possibilities for a hand of five cards. Thus, the probability is approximately  $1.5 \cdot 10^{-6}$ , very low indeed.  
**STRAIGHT FLUSH:** The highest card of the straight must be between 5 and K, i. e., there are 9 possibilities for each suit. Altogether, the probability is  $\frac{36}{2598960} \doteq 1.4 \cdot 10^{-5}$ .  
**POKER (FOUR OF A KIND):** There are 13 possibilities

It is apparent that if the matrix is represented by an array of zeros and ones, then the first and second operations can be executed in  $O(1)$  (constant time), while the others are in  $O(n)$  (linear time).

In the case of the adjacency list, implementation of the data structures is crucial for the time complexity. However, all of the operations should be proportional to the number of edited data units provided the corresponding item(s) are already found. For instance, if a vertex is removed, then all of the edges that are incident to it must also be removed.

The matrix representation is also useful in theoretical discussions about graphs, using matrix calculus.

**13.2.2. Searching in graphs.** Many useful algorithms are based on going through all vertices of a given graph step by step. Usually, the vertex is given to start with or it is selected at the beginning of the procedure.

At every stage of the search, each vertex has (exactly) one of the following statuses:

- *processed* – it has been visited and completely processed;
- *active* – it has been visited and is prepared to be processed;
- *sleeping* – it has not been visited yet.

At the same time, information about processed edges is retained. At every stage, the sets of vertices and/or edges in these groups must form a partition of the sets  $V$  and  $E$  while one of the active vertices is being processed.

The general principle on searching through the vertices is illustrated first. In the subsequent subsections, such procedures are used to build algorithms solving particular problems.

At the beginning of the algorithm, there is just one active vertex and all the others are sleeping. At the first step, traverse all edges incident to the active vertex and change the status of their other endpoints from sleeping to active. Then, the active vertex started may be marked as processed, and another active vertex may be chosen. In the following steps, always go through those adjacent edges not yet met, marking their other endpoints as active. This algorithm can be applied to both directed and undirected graphs.

In practical problems, the search is often restricted to only some edges going from the current vertex. This is an insignificant change to the algorithm.

To specify the algorithm completely, a decision must be made in which order to process the active vertices and in which order to process the edges going from the current vertex. In general, the two simplest possibilities of processing the vertices are:

- (1) they are processed in the same order as they were visited (queue),
- (2) they are processed in the reversed order than they were visited (stack).

The former case, is called a *breadth-first search*. The latter case is called a *depth-first search*.

for the quad and the fifth card can be arbitrary (48 possibilities). Hence:  $\frac{624}{2598960} \doteq 2.4 \cdot 10^{-4}$ .

FULL HOUSE: There are  $13 \binom{4}{3} = 52$  possibilities for the triple and  $12 \binom{4}{2} = 72$  possibilities for the remaining pair. Altogether,  $\frac{3744}{2598960} \doteq 1.4 \cdot 10^{-3}$ .

FLUSH: There are 4 suits and  $\binom{13}{5}$  hands for each suit, i. e.,  $4 \cdot \binom{13}{5} = 5148$  possibilities in total. However, we must not count the straights again. There are 40 of them, so the resulting probability is  $\frac{5108}{2598960} \doteq 2 \cdot 10^{-3}$ .

STRAIGHT: The highest card of the straight is between 5 and A, so there are 10 possibilities. Selecting the suit of each card arbitrarily, this gives  $10 \cdot 4^5 = 10240$  possibilities. However, we must exclude flushes, so the total probability is  $\frac{10200}{2598960} \doteq 3.9 \cdot 10^{-3}$ .

Altogether, the probability of one of the best six combinations is approximately  $3.9 \cdot 10^{-3} + 2 \cdot 10^{-3} + 1.4 \cdot 10^{-3} + 0.24 \cdot 10^{-3} = 7.54 \cdot 10^{-3}$ , i. e., about 0.75%.

In the Texas Hold 'em variation, the best 5-card hand of the seven cards is always considered. We have computed the number of favorable 5-card hands and there are  $\binom{52-5}{2}$  possibilities for the remaining two cards. The total number of 7-card hands is  $\binom{52}{7}$ . We can thus approximate the probability for Texas Hold 'em for the classic Poker by multiplying by the coefficient  $\frac{\binom{52}{5} \binom{47}{2}}{\binom{52}{7}} = 21$ .

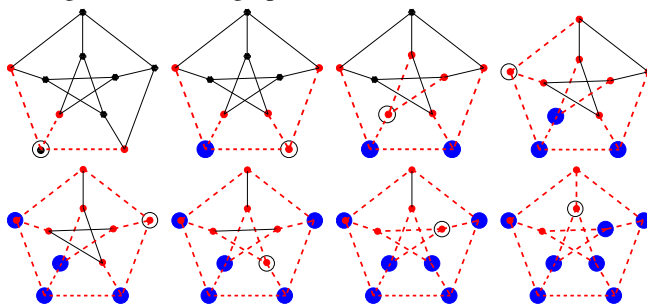
However, note that this is indeed only an approximation of the actual probability since some favorable combinations are counted more than once this way. For instance, we have a full house in the considered 5-card hand and if the arbitrary pair contains the fourth card of the triple, then we actually have a poker (four of a kind), so this combination has been counted more times. On the other hand, the true result only merely differs from the computed approximation, so the probability of one of the best six poker combinations is about twenty times higher than in classic Poker. This may be the reason why this variation is so popular.

- iv) Clearly, our situation is very good. Hence, it will be easier to count the unfavorable cases when the other player has a better combination. We have Twos full of Aces, so we lose only if the opponent has Aces full of Twos or a poker of Twos, i. e., he must hold the remaining Two or an Ace. In the former case, he surely wins, and this happens in  $0 + 3 + 4 + \dots + 4 + 2 = 45$  cases (we can see the remaining Twos, a Three, and two Aces) out of all  $\binom{46}{2} = 1035$ , so the probability of this loss is about 0.043. In the latter case, there are more possibilities. If he holds a pair of Aces and the river card is not the remaining Two, then we lose; otherwise (i. e., if he has only one Ace or the Two appears on the river), it is a tie. Thus, we lose in this case by  $\frac{1}{1035} \cdot \frac{43}{44} \doteq 10^{-3}$ . Altogether, the probability that we win or draw is almost 96 %.

The role of the data structures used for representing the graph is immediately apparent: The adjacency list allows passage through all edges going from a given vertex in a time proportional to the number of them. Each edge is visited at most twice since it has only two endpoints. Hence the following result:

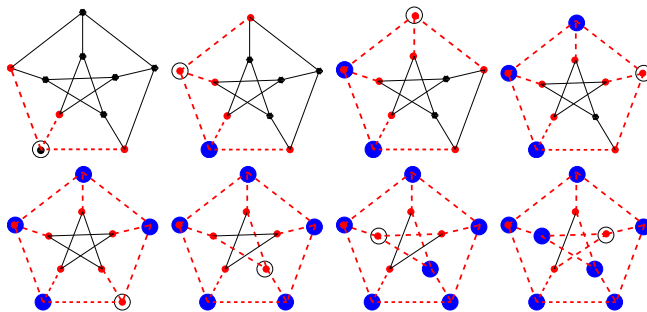
**Theorem.** Both the breadth-first and depth-first searches run in  $O((n + m)K)$  time, where  $n$  and  $m$  are the number of vertices and edges of the graph, respectively.  $K$  is the time needed for processing an edge or a vertex.

The following diagram illustrates the breadth-first search through the Petersen graph:



The first 8 steps are shown here. The circled vertex is the one to be processed, the bold vertices are the already processed one, while the dashed edges are those that have been processed, and the small vertices adjacent to some dashed edges are the active ones. At the given vertex, the edges are processed counterclockwise, beginning with the direction “straight down”.

The diagram below illustrates the depth-first search applied to the same graph. Note that the first step is the same as above.



As a simple example of graph searching, consider an algorithm for finding all connected components of a given graph. The only information that must be processed during the search (no matter whether breadth-first or depth-first) is which component is being examined.

The search, as here presented, passes exactly the vertices of a single component. Hence, one can start with all vertices in the sleeping state and choose any one of them. During the search, whenever there are no more active vertices to be processed, the search of one component is finished. One can then choose an arbitrary sleeping vertex and continue likewise. The algorithm terminates as soon as there are no more sleeping vertices remaining.

**13.E.11.** Four players are given two cards each from a pack consisting of four Aces and four Kings. What is the probability that at least one of the players is given a pair of Aces? Express the result as a ratio of two-digit integers. ◻

**13.E.12.** Alex owns two special dice: one of them has only 6's on its sides. The other one has two 4's, two 5's, and two 6's. Martin has two ordinary dice. Each of the players throws his dice and the one whose sum is higher wins. What is the probability that Alex wins? Express the result as a ratio of two-digit integers. ◻

**13.E.13.** In how many ways can we place  $n$  rooks on an  $n \times n$  chessboard so that every unoccupied square is guarded by at least one of the rooks?

**Solution.** Clearly, the condition is satisfied if and only if at least one of the following holds: There is at least one rook in each rank (which implies that there must be exactly one rook in each rank—there are  $n^n$  such placements since the particular squares can be selected independently for each rank); or there is at least one rook in each file (again resulting in  $n^n$  placements). However, there are  $n!$  placements which satisfy both (we have  $n$  squares where to put a rook in the first rank,  $n - 1$  squares for the second rank since one of the files is already occupied, etc.). By the inclusion-exclusion principle, the wanted number of placements is:

$$2n^n - n!. \quad \square$$

**13.E.14.** We flip a coin five times. Every time it comes up heads, we put a white ball into a hat. Every time it comes up tails, we put a black ball into the hat. Express the probability that there are more black balls than white ones provided there is at least one black ball.

**Solution.** Let us define the events

- $A$  – there are more black balls than white ones,
- $H$  – there is at least one black ball.

We want to compute  $P(A|H)$ . Clearly, the probability  $P(H^C)$  of the complementary event to  $H$  is  $2^{-5}$ . Further, the probability of  $A$  is the same as the probability  $P(A^C)$ . Therefore,  $P(H) = 1 - 2^{-5}$  and  $P(A) = 1/2$ . Further,  $P(A \cap H) = P(A)$  since the event  $H$  contains the event  $A$  (the event  $A$  implies the event  $H$ ). Altogether, we have obtained

$$P(A|H) = \frac{P(A \cap H)}{P(H)} = \frac{\frac{1}{2}}{1 - (\frac{1}{2})^5} = \frac{16}{31}. \quad \square$$

**F. More advanced problems from combinatorics**

In the first chapter, we met the fundamental methods used in combinatorics. Even using merely these ideas, we are able to solve relatively complicated problems.

◻ **13.2.3. Natural metrics on graphs.** The concept of “path length” is used earlier. This recalls the general idea of distance. The concept of distance in graphs can be built mathematically in this manner.



For an (undirected) graph, define the *distance between vertices*  $v$  and  $w$  to be the number  $d_G(v, w)$ . This is the number of edges on the shortest path from  $v$  to  $w$ . If there is no such path, write  $d_G(v, w) = \infty$ .

For the sake of simplicity, consider only connected graphs  $G$ . The function  $d_G : V \times V \rightarrow \mathbb{N}$  defined as above satisfies the usual three properties of a distance (it is recommended to compare this to the issues from the relevant part of chapter seven, see 7.3.1 (the page 666):

- $d_G(v, w) \geq 0$ , and  $d_G(v, w) = 0$  if and only if  $v = w$ ;
- the distance is symmetric, i.e.  $d_G(v, w) = d_G(w, v)$ ;
- the triangle inequality holds; i.e. for every triple of vertices  $v, w, z$ ,

$$d_G(v, z) \leq d_G(v, w) + d_G(w, z).$$

$d_G$  is a *metric on the graph*  $G$ .

Besides these three properties, every such metric on a graph apparently satisfies the following:

- $d_G(v, w)$  is always a non-negative integer;
- if  $d_G(v, w) > 1$ , then there exists a vertex  $z$  distinct from  $v$  and  $w$  such that  $d_G(v, w) = d_G(v, z) + d_G(z, w)$ .

The following is true:



Every function  $d_G$  on  $V \times V$  (for a finite set  $V$ ), satisfying the five properties listed above, allows to define the edges  $E$  so that  $G = (V, E)$  is a graph with metric  $d_G$ .

Prove this yourself as an exercise! (It is quite clear how to construct the corresponding graph. It remains “merely” to show that the given function  $d_G$  could be achieved as the metric on the constructed graph.)

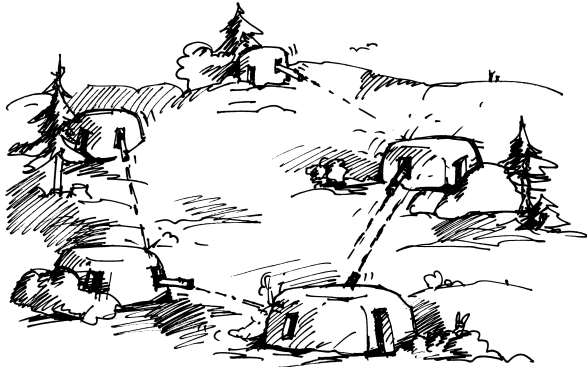
**13.2.4. Dijkstra’s shortest-path algorithm.** One may suspect that the shortest path between a given vertex  $v$  and another given vertex  $w$  can be found by breadth-first searching the graph. With this approach, discuss first the vertices which are reachable with one edge from the initial vertex  $v$ , then those which are two edges distant, and so on. This is the fundamental idea of one of the most often used graph algorithms – the *Dijkstra’s algorithm*<sup>5</sup>.



This algorithm is able to find the shortest paths even in problems from practice, where each edge  $e$  is assigned a *weight*  $w(e)$ , which is a positive real number. When looking for shortest paths, the weights are to represent lengths of the

<sup>5</sup>Edsger Wybe Dijkstra (1930 - 2002) was a famous Dutch computer scientist, being one of the fathers of this discipline. Among others, he is credited as one of founders of concurrent computing. He published the above algorithm in 1959

**13.F.1.** There are  $n$  ( $n \geq 3$ ) fortresses positioned on a circle, numbered 1 through  $n$ . At a given moment, every fortress shoots at one of its neighbors (i. e., fortress 1 shoots at  $n$  or 2, fortress 2 shoots at 1 or 3, etc.). We will refer to the set of hit fortresses as a *result* (i. e., we are only interested in whether each fortress was hit or not; it does not matter whether it was hit once or twice). Let  $P(n)$  denote the number of possible results. Prove that the integers  $P(n)$  and  $P(n + 1)$  are coprime.



**Solution.** First of all, note that a set of hit fortresses is a possible result if and only if no pair of adjacent-but-one fortresses (i. e., whose numbers differ by 2 modulo  $n$ ) is unhit. Therefore, if  $n$  is odd, then  $P(n)$  is equal to the number  $K(n)$  of results where no pair of *adjacent* fortresses is unhit (consider the order  $1, 3, 5, \dots, n, 2, 4, \dots, n - 1$ ). If  $n$  is even, then  $P(n)$  equals  $K(n/2)^2$  since fortresses at even positions and those at odd positions can be considered independently.

We can easily derive the following recurrent formula for  $K(n)$ :  $K(n) = K(n - 1) + K(n - 2)$ . (Well, on the other hand, it is not so trivial... It is left as an exercise for the reader.) Further, we can easily calculate that  $K(2) = 3$ ,  $K(3) = 4$ ,  $K(4) = 7$ , so  $K(2) = F(4) - F(0)$ ,  $K(3) = F(5) - F(1)$ ,  $K(4) = F(6) - F(2)$ , and simple induction argument shows that  $K(n) = F(n + 2) - F(n - 2)$ , where  $F(n)$  denotes the  $n$ -th term of the Fibonacci sequence ( $F(0) = 0$ ,  $F(1) = F(2) = 1$ ). Moreover, since  $(K(2), K(3)) = 1$ , we have for  $n \geq 3$  that (similarly as with the Fibonacci sequence)

$$\begin{aligned} (K(n), K(n - 1)) &= (K(n) - K(n - 1), K(n - 1)) = \\ &= (K(n - 2), K(n - 1)) = \dots = 1. \end{aligned}$$

Now, we are going to show that, for every  $n = 2a$ ,  $P(n) = K(a)^2$  is coprime to both  $P(n + 1) = K(2a + 1)$  and  $P(n - 1) = K(2a - 1)$ . It suffices to realize that for  $a \geq 2$ , we have

$$\begin{aligned} (K(a), K(2a + 1)) &= (K(a), F(2)K(2a) + F(1)K(2a - 1)) \\ &= (K(a), F(3)K(2a - 1) + F(2)K(2a - 2)) = \dots \\ &= (K(a), F(a + 1)K(a + 1) + F(a)K(a)) \\ &= (K(a), F(a + 1)) = (F(a + 2) - F(a - 2), F(a + 1)) \\ &= (F(a + 2) - F(a + 1) - F(a - 2), F(a + 1)) \\ &= (F(a) - F(a - 2), F(a + 1)) \end{aligned}$$

edges. However, in general, the weights may have other meanings: they may stand for profits or costs, network flows, and so on.

The input of the algorithm consists of an edge-weighted graph  $G = (V, E)$  and an initial vertex  $v_0$ . The output consists of the numbers  $d_w(v)$ , which give the least possible sum of the weights of edges along a path from the vertex  $v_0$  to the vertex  $v$ . This procedure works in undirected graphs as well as in directed ones.

In order to ensure the termination and the correctness of the algorithm, it is important that all of the weights are positive – see example 13.B.6. Dijkstra's algorithm needs only a little modification of the general breadth-first search:



- For every vertex  $v$ , keep the information  $d(v)$ , which is an upper bound for the actual distance of  $v$  from the initial vertex  $v_0$ .
- At every stage, the already processed vertices are those for which the shortest path is already known. Then,  $d(v) = d_w(v)$ .
- When some sleeping vertices are to be made active, choose exactly those vertices  $y$  from the set  $Z$  of sleeping vertices for which  $d(y) = \min\{d(z); z \in Z\}$ .

Suppose that the graph  $G$  has at least two vertices. More formally, Dijkstra's algorithm can be described as follows:

#### DIJKSTRA'S ALGORITHM

*Input:* vertex  $v_0$  in the graph  $G = (V, E)$  with weights on all edges.

*Output:* the distances from  $v_0$  within  $G$  associated to all vertices.

- (1) *Initialization step:* Set the values for all  $v \in V$ :

$$d(v) = \begin{cases} 0 & \text{for } v = v_0, \\ \infty & \text{for } v \neq v_0. \end{cases}$$

Set  $Z = V$ ,  $W = \emptyset$ .

- (2) *Cycle condition:* If every vertex  $y \in Z$  is assigned  $\infty$ , the algorithm terminates; otherwise the algorithm continues with another iteration. (In particular, the algorithm terminates if  $Z = \emptyset$ .)

- (3) *Update of the vertex statuses:*

- Find the set  $N$  of those vertices  $v \in Z$  for which  $d(v) = \delta$  is as small as possible:

$$\delta = \min\{d(y); y \in Z\}.$$

- All vertices which have been in  $W$  are removed and marked as processed; the new set of active vertices is  $W = N$ , while all these vertices are removed from  $Z$ , i.e. they are no more sleeping.

- (4) *Cycle body:* For each edge  $e \in E_{WZ}$  (i.e. whose tail is an active vertex  $v$  and head is a sleeping vertex  $y$ :

- if  $d(v) + w(e) < d(y)$ , then update  $d(y)$  to  $d(v) + w(e)$ .

Move back to check the cycle condition (step 2).

$$\begin{aligned}
 &= (F(a-1), F(a+1)) = (F(a-1), F(a)) = 1. \\
 &(K(a), K(2a-1)) \\
 &= (K(a), F(2)K(2a-2) + F(1)K(2a-3)) \\
 &= (K(a), F(3)K(2a-3) + F(2)K(2a-4)) \\
 &= \dots = (K(a), F(a)K(a) + F(a-1)K(a-1)) \\
 &= (K(a), F(a-1)) = (F(a+2) - F(a-2), F(a-1)) \\
 &= (F(a+2) - F(a), F(a-1)) \\
 &= (F(a+2) - F(a+1), F(a-1)) = (F(a), F(a-1))
 \end{aligned}$$

This proves the proposition.  $\square$

### G. Probability in combinatorics

Classical probability is tightly connected to combinatorics, as we have already seen in the first chapter. Now, we present another example, which is a bit more complicated.

Combinatorics is hidden even in the following “probabilistic” problem.

**13.G.1.** There are 100 prisoners in a prison, numbered 1 through 100. The chief guard has placed 100 chests (also numbered 1 through 100) into a closed room and randomly put balls with numbers 1 through 100 on them into the chests so that each chest contains exactly one ball. He has decided to play the following game with the prisoners: He calls them one by one into the room and the invited prisoner is allowed to gradually open 50 chests. Then, he leaves without any possibility to talk to the other prisoners, the guard closes all the chests, and another prisoner is let in. The guard has promised to free all the prisoners provided each of them finds the ball with his number in one of the 50 opened chests. However, if any of the prisoners fails to find his ball, all will be executed. Before the game begins, the prisoners are allowed to agree on a strategy. Does there exist a strategy that gives the prisoners a “reasonable” chance of winning?



**Solution.** Clearly, if the prisoners choose to open the chests randomly (where the choices of the particular prisoners are independent), the chance for one prisoner to find his ball is  $1/2$ , so the total probability of success is merely  $1/2^{100}$ . Therefore, it is necessary to look for a strategy where the successes of the prisoners are as dependent as possible. First of all, we should realize that the invited prisoner has no information from other prisoners and does not know the positions of particular balls in the chests. However, once he opens a chest, he knows the ball number it contains and may choose the next chest accordingly. This suggests the following simple strategy: Every prisoner starts with the chest that bears his number. If it contains the corresponding ball, the prisoner succeeds and can open the remaining chests at random. If not, he opens the chest with the number of the found ball. He continues this way until he eventually finds his ball or opens the fiftieth chest. Since every chest “points” at another chest according to the described procedure, let us call this strategy the *pointer strategy*.

**13.2.5. Theorem.** For a given vertex  $v_0$ , the Dijkstra’s algorithm finds the distance  $d_w(v)$  of each vertex  $v$  in  $G$  that lies in the connected component of the vertex  $v_0$ . For the vertices  $v$  of other connected components,  $d(v) = \infty$  remains.

The algorithm can be implemented in such a way that it terminates in time  $O(n \log n + m)$ , where  $n$  is the number of vertices and  $m$  is the number of edges in  $G$ .

**PROOF.** The algorithm is correct, since

- it terminates after a finite number of steps;
- when it does, its output has the desired properties.

The cycle condition guarantees that in each iteration, the number of sleeping vertices decreases by one at least since  $N$  is always non-empty. Therefore, the algorithm necessarily terminates after a finite number of steps.

After going through the initialization cycle,

$$(1) \quad d_w(v) \leq d(v)$$

for all vertices  $v$  of the graph. Now assume that this property holds when the algorithm enters the main cycle and show that it holds when it leaves the cycle as well. Indeed, if  $d(y)$  is changed during step 4, then it is caused by finding a vertex  $v$  such that

$$d_w(y) \leq d_w(v) + w(\{v, y\}) \leq d(v) + w(\{v, y\}) = d(y),$$

where the new value is on the right-hand side.

The inequality (1) is satisfied when the algorithm terminates. It remains to verify that the other inequality holds as well. For this purpose, consider what is actually done in steps 3 and 4 of the algorithm.

Let  $0 = d_0 < \dots < d_k$  denote all (distinct) finite distances  $d_w(v)$  of the vertices in  $G$  from the initial vertex  $v_0$ . At the same time, this partitions the vertex set of the graph  $G$  into clusters  $V_i$  of vertices whose distance from  $v_0$  is exactly  $d_i$ . During the first iteration of the main cycle,  $N = V_0 = \{v_0\}$ , the number  $\delta$  is just  $d_1$ , and the set of sleeping vertices is changed to  $V \setminus V_0$ .

Suppose this holds up to  $j$ -th iteration (inclusive), i.e. the algorithm enters the cycle with  $N = V_j$ ,  $\delta = d_j$ , and  $\bigcup_{i=0}^j V_i = V \setminus N$ . Consider a vertex  $y \in V_{j+1}$ , i.e.  $d_w(y) = d_{j+1} < \infty$ , and there exists a path  $(v_0, e_1, v_1, \dots, v_\ell, e_{\ell+1}, y)$  with total length  $d_{j+1}$ . However, then

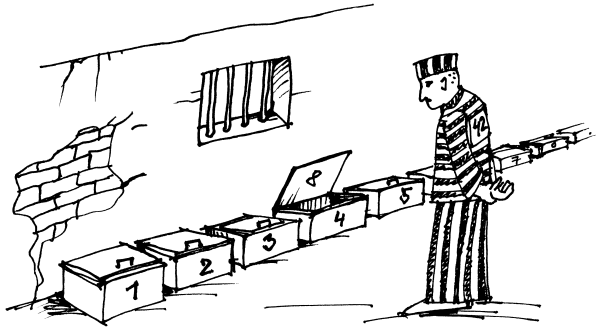
$$(2) \quad d_w(v_\ell) \leq d_{j+1} - w(\{v_\ell, y\}) < d_{j+1}.$$

It follows from the assumption that the vertex  $v_\ell$  was active during an earlier iteration of the main cycle, and  $d_w(v_\ell) = d(v_\ell) = d_i$  for some  $i \leq j$  then. Therefore, after the current iteration of the main cycle has been finished,

$$d(y) = d_w(v_\ell) + w(\{v_\ell, y\}) = d_{j+1}$$

and this does not change any more. It follows that the inequality (1) holds with equality when the algorithm terminates.





**Probability of success.** The guard's possible placements of the balls bijectively correspond to permutations of the numbers 1 through 100. In order to find the probability of success, we must realize for which permutations the pointer strategy works. Recall that every permutation can be expressed as the composition of pairwise disjoint cycles. If each prisoner were allowed to open an arbitrary number of chests, he would find his ball as the last one of the corresponding cycle since he begins with the chest with his number, which is pointed at just by the chest with his ball. It follows that the strategy fails if and only if there is a cycle of length greater than 50 because then no prisoner of this cycle finds his ball in time. Thus, we must count the number of such permutations. In general, the probability that a random permutation of length  $n$  contains a cycle of length  $r > n/2$  (there could be more occurrences of shorter cycles; however, there can be at most one cycle of length greater than  $n/2$ , which simplifies the calculation) is as follows: We must choose the  $r$  elements of the cycle, order them, and then choose an arbitrary permutation of the remaining  $n - r$  numbers. This leads to

$$\binom{n}{r} (r - 1)! (n - r)! = \frac{n!}{r}$$

Therefore, the probability that such permutation is selected (among all the  $n!$  permutations) is  $1/r$ . Thus, the probability that our 100 prisoners succeed is

$$1 - \sum_{k=51}^{100} \frac{1}{k} \approx 0.311828$$

As we can see, this is much higher than the original  $1/2^{100}$ . Now, let us look at the behavior of this probability for a general number  $n$  of the prisoners (then, each prisoner is allowed to open at most  $n/2$  chests). In general, the probability that a random permutation of length  $n$  contains a cycle of length greater than  $n/2$  is equal to

$$p = \sum_{k=1+\frac{n}{2}}^n \frac{1}{k}$$

Recall that  $\sum_{k=1}^n \frac{1}{k} \rightarrow \ln(n) + \gamma$  for  $n \rightarrow \infty$ , where  $\gamma$  is Euler's constant. Thus, we have:

The analysis of the main cycle just made also determines a bound for the running time of the algorithm (i.e. the number of elementary operations with the graph and other corresponding objects). The main cycle is iterated as many times as there are (distinct) distances  $d_i$  in the graph. Every vertex, when processed during step 3, is considered exactly once. The vertices that are still sleeping must be sorted. This gives the bound  $O(n \log n)$  for this part of the algorithm provided the graph is stored as a list of vertices and weighted edges, such that the sleeping vertices are kept in a suitable data structure that allows the finding of the set of  $N$  active vertices in time  $O(\log n + |N|)$ . This can be achieved if a heap is used. Every edge is processed exactly once in step 4 since the vertices are active only during one iteration of the cycle.  $\square$

Note that the inequality (2), essential for the analysis of the algorithm, need not hold if the weights of the edges are allowed to be negative.

In practice, many heuristic improvements of the algorithm are applied. For instance, it is not necessary to compute the distance between all vertices if only the distance between a given pair of vertices is of interest. When the vertex is excluded from the active ones, its distance is final.



Further, it is not necessary to initialize the distances with the value of infinity. Of course, this is technically impossible, and a sufficiently large constant would be needed in the implementation.

**13.2.6. Spanning trees.** In practical applications, graphs often encode all possibilities of connections between particular objects, as in road or electrical networks. If it is only required that each pair of vertices is connected by a path, using as few edges as possible, then what is needed is a subgraph  $T$  which is a tree. This corresponds to the problem of finding a type of minimal network.



#### SPANNING TREE OF A GRAPH

**Definition.** Any tree  $T = (V, E')$  in a graph  $G = (V, E)$ ,  $E' \subseteq E$  is called a *spanning tree* of the graph  $G$ .

A graph can have a spanning tree if and only if it is connected.

A spanning tree is connected since all trees are. Conversely, the following algorithm finds a spanning tree for any given connected graph.

$$p = \sum_{k=1}^n \frac{1}{k} - \sum_{k=1}^{\frac{n}{2}} \frac{1}{k} \rightarrow \ln(n) + \gamma - \ln\left(\frac{n}{2}\right) - \gamma = \ln 2, \text{ pro } n \rightarrow \infty$$

Hence it follows that, for large values of  $n$ , the probability of success approaches  $1 - p \simeq 1 - \ln 2 = 0.30685 \dots$ . Now, we are going to show that the pointer strategy is optimal. **Optimality of the pointer strategy.** In order to prove the optimality of the pointer strategy, we merely modify the rules of the game and further define another game.

Consider the following rules: Every prisoner keeps opening the chests until he finds his ball. The prisoners win iff each opens at most 50 chests. Clearly, this modification does not change the probability of success, but it will help us prove the optimality. We will refer to this game as game A.

Now, consider another game (game B) with the following rules: First, prisoner number 1 enters the chest room and keeps opening the chests until he finds his ball. Then, the guard leaves the opened chests as they are and immediately invites the prisoner with the least undiscovered number. The game proceeds this way until all chests are opened. The prisoners win iff none of them opened more than 50 ( $n/2$  in the general case) chests.

Suppose that the guard notes the ball numbers in the order they were discovered by the prisoners. This results in a permutation of the numbers 1 through 100, from which he can see whether the prisoners won or not. The probability of discovering a particular ball is at every moment independent of the selected strategy. There are 100! permutations which correspond to some strategies (no matter whether they are random or sophisticated) since they are merely the orders in which the ball numbers are discovered. In order to compute the probability of success in game B, we should note that any order can be written as the composition of cycles where each cycle contains the ball numbers discovered by a given prisoner. For the sake of clarity, consider a game with 8 prisoners. If the guard has noted the permutation (2, 5, 7, 1, 6, 8, 3, 4), then we can see that the prisoners win since prisoner 1 discovered numbers (2, 5, 7, 1), then prisoner 3 discovered (6, 8, 3), and finally prisoner 4 discovered only his number (4). In this case, we can write: (2, 5, 7, 1, 6, 8, 3, 4)  $\rightarrow$  (2, 5, 7, 1)(6, 8, 3)(4). Further, any such permutation corresponds to a unique order of the numbers 1 through 8. Having any permutation in the cyclic notation, we first rearrange each cycle so that its least number is the last one and then sort the cycles by their last numbers in ascending order. For instance, we have:

$$(7, 5, 8)(2, 4)(1, 6, 3) \rightarrow (6, 3, 1)(4, 2)(8, 7, 5) \rightarrow (6, 3, 1, 4, 2, 8, 7, 5).$$

We have thus constructed a bijection between the winning orders of discovered numbers and the permutations of the numbers 1 through 8 that do not contain a cycle of length greater than 4. It follows that the probability of success in game B is the same as the probability that a random permutation does

SPANNING FOREST ALGORITHM

**Input:** Graph  $G = (V, E)$   
**Output:** A forest  $T = (V, E')$  consisting of spanning trees of the components of  $G$ .

- (1) Sort all edges  $e_0, \dots, e_m \in E$  in any order.
- (2) Start with  $E_0 = \{e_0\}$  and gradually build the sets of edges  $E_i$  so that in the  $i$ -th step add the edge  $e_i$  to  $E_{i-1}$  unless this creates a cycle in the graph  $G_i = (V, E_{i-1} \cup \{e_i\})$ . If this edge creates a cycle, leave  $E_i = E_{i-1}$  unchanged.
- (3) The algorithm terminates if the graph  $G_i = (V, E_i)$  has exactly  $n - 1$  edges at some step  $i$  or if  $i = m$ , and produces the graph  $T = (V, E_i)$ .

If the algorithm terminates for the latter reason, then the graph is not connected and no spanning tree exists (but there are still the spanning trees of all individual components).

**PROOF.** It follows from the rules of the algorithm that the resulting subgraph  $T$  of  $G$  never contains a cycle. Therefore, it is a forest. If the resulting number of edges is  $n - 1$ , then it must be a tree; see theorem 13.1.15.

It remains to show that the connected components of the graph  $T$  have the same sets of vertices as the connected components of the original graph  $G$ : Every path in  $T$  is also a path in  $G$ ; therefore, all vertices that lie in one tree of  $T$  must lie in the same component of  $G$ . If there is a path in  $G$  from  $v$  to  $w$  such that its endpoints lie in different trees of  $T$ , then one of its vertices  $v_i$  is the last one that is in the component determined by  $v$  (in particular,  $v_{i+1}$  does not lie in this component). The corresponding edge  $\{v_i, v_{i+1}\}$  creates a cycle when examined by the algorithm since otherwise, it would be in  $T$ . Since the edges are never removed from  $T$ , there is a path between  $v_i$  and  $v_{i+1}$  in  $T$ , which contradicts the assumptions. Therefore,  $v$  and  $w$  cannot lie in different trees of  $T$ . The number of components of  $T$  is given by the fact that the number of vertices and edges differs by one in every tree. The difference increases by one with every component so if there are  $n$  vertices and  $k$  edges in the forest, then there are  $n - k$  components.  $\square$

**Remark.** As always, the time complexity of the algorithm is of interest. The addition of an edge creates a cycle if and only if its endpoints lie in the same connected component of the forest  $T$  under construction.

Knowledge of the connected components of the current forest  $T$  is helpful. To implement the algorithm, it is needed to unite two equivalence classes of a given equivalence relation on a given set (the vertex set) and to find out whether two vertices are in the same class or not. The union requires time  $O(k)$ , where  $k$  is the number of elements to be united.  $k$  can be bounded from above by  $n$ , the total number of vertices.

However, for each equivalence class it can be noted how many vertices it contains. If, for each vertex, the information to which class it belongs is kept, then the union operation



not contain a cycle of length greater than 4 ( $n/2$  in the general case). This corresponds to the probability of success in the original game using the pointer strategy. Now, this implies an important conclusion for game A. Indeed, the prisoners may apply any strategy from game A to game B as follows: each prisoner behaves like in game A, but he considers open chests to be closed, i. e., if he wants to open a chest which has already been opened, he just “p”asses this move and further behaves as if he had just discovered the ball number in the considered chest. Therefore, any strategy that succeeds for a given placement of the balls in game A must succeed for the same placement in game B as well. Therefore, if there existed a better strategy for game A, we could apply it to game B and obtain a higher chance of winning there. However, this is impossible since all strategies in game B lead to the same probability of success. Therefore, the pointer strategy is better than or equally good as any other strategy.  $\square$

**13.G.2.** In a competition, there are  $m$  contestants and  $n$  officials, where  $n$  is an odd integer greater than two. Each official judges each contestant as either good or bad. Suppose that any pair of officials agree on at most  $k$  contestants. Prove that

$$\frac{k}{m} \geq \frac{n-1}{2n}.$$

Let us look at two possible approaches to this problem.

**Solution.** Let us count the number  $N$  of pairs ( $\{\text{official, official}\}$ , contestant) where the officials are distinct and agree on the contestant. Altogether, there are  $\binom{n}{2}$  pairs of officials, and each pair can agree on at most  $k$  contestants. Therefore,  $N \leq k\binom{n}{2}$ .

Now, let us fix a contestant  $X$  and count the number of pairs of officials who agree on  $X$ . Say that  $x$  officials said  $X$  was good. Then, there are  $\binom{x}{2}$  pairs who agree that  $X$  is good and  $\binom{n-x}{2}$  pairs who agree that  $X$  is bad. Altogether, there are

$$\binom{x}{2} + \binom{n-x}{2} = \frac{x(x-1)}{2} + \frac{(n-x)(n-x-1)}{2}$$

pairs that agree on  $X$ . We have:

$$\begin{aligned} \frac{x(x-1)}{2} + \frac{(n-x)(n-x-1)}{2} &= \frac{2x^2 - 2nx + n^2 - n}{2} = \\ &= \left(x - \frac{n}{2}\right)^2 + \frac{n^2}{4} - \frac{n}{2} \geq \frac{n^2}{4} - \frac{n}{2} = \frac{(n-1)^2}{4} - \frac{1}{4}. \end{aligned}$$

Since  $n$  is odd, the expression  $(n-1)^2/4$  is an integer. Thus, the number of pairs that agree on  $X$  is at least  $(n-1)^2/4$ . Hence  $N \geq m(n-1)^2/4$ . Combining these two inequalities together, we get

$$\frac{k}{m} \geq \frac{n-1}{2n}.$$

**An alternative solution - using probabilities.** Let choose a pair of officials at random. Let  $X$  be the random variable which tells the number of cases when this pair agrees. We are going to prove the contrapositive implication, i. e., if  $\frac{k}{m} < \frac{n-1}{2n}$ , then  $X$  is greater than  $k$  with probability greater than zero, which will be denoted  $P(X > k) > 0$ .

means to relabel the vertices of one of the united classes. If the smaller class is always selected to be relabeled, then the total number of operations of the algorithm is  $O(n \log n+m)$ . (As an exercise, complete the details of these considerations yourself!)



The above reasoning shows that slightly better in time might be achieved, if only the spanning tree of the connected component of a given starting vertex is of interest:

ANOTHER SPANNING TREE ALGORITHM

**Input:**  $G = (V, E)$  with  $n$  vertices and  $m$  edges, vertex  $v \in V$ .

**Output:** The tree  $T$  spanning the connected component of  $v$ .

- (1) Initialize  $T_0 = (\{v\}, \emptyset)$ .
- (2) In the  $i$ -th step, we build the tree  $T_i$  as follows. Look for edges  $e$  which are not in  $T_{i-1}$ , but their tail vertex  $v_e$  is. Take one of them and add it to  $T_{i-1}$ , i.e. add the head vertex to  $V_{i-1}$  and  $e$  to  $E_{i-1}$ .
- (3) The algorithm terminates as soon as no such edge exists. Apparently, the resulting graph  $T$  is connected. The count of its vertices and edges shows that it is a tree.

**PROOF.** The vertices of  $T$  coincide with the vertices of the connected component of the graph  $G$  containing the starting vertex  $v$ .

Suppose there is a path from  $v$  to a given vertex  $w$ . If  $w$  does not lie in  $T$ , then label by  $v_i$  the last of its vertices that lie in  $T$  (just like in the proof of the previous lemma). However, the subsequent edge of the path would have to be added to  $T$  by the algorithm when it terminated, which is a contradiction.

Consequently, this algorithm finds a spanning tree of the connected component that contains a given initial vertex  $v$  in time  $O(n+m)$ .  $\square$

**13.2.7. Minimum spanning tree.** All spanning trees of a given graph  $G$  have the same number of edges since this is a general property of trees. Just as



the shortest path in graphs with weighted edges was found, now spanning trees with the minimum sum of their edges' weights is desired.

**Definition.** Let  $G = (V, E, w)$  be a connected graph whose edges  $e$  are labeled by non-negative weights  $w(e)$ . A *minimum spanning tree* of  $G$  is such that its total weight does not exceed that of any other spanning tree.

This problem has many applications in practice. For instance, networks of electricity, gas, water, etc.

Surprisingly, it is quite simple to find a minimum spanning tree (supposing all edge weights  $w(e)$  of  $G$  are non-negative) by the following procedure<sup>6</sup>:

<sup>6</sup>Joseph Bernard Kruskal (1928 - 2010) was a famous American mathematician, statistician, computer scientist, and psychometrician. There are other famous mathematicians of the same surname - his two brothers and one nephew. Martin David co-invented solitons and surreal numbers, William

Consider the random variables  $X_i$  for  $i = 1, 2, \dots, m$  with codomain  $0, 1$ , denoting whether the pair agrees on the  $i$ -th contestant. Let  $X_i = 1$  when they agree, and let  $X_i = 0$  otherwise. Hence we have:

$$X = X_1 + X_2 + \dots + X_m$$

Using the linearity of expectation, we obtain:

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_m].$$

Now, let us calculate  $E[X_i] = \sum_{x_i \in \{0,1\}} x_i \cdot P(X_i = x_i)$ . Since  $X_i$  can be only 0 or 1, we have directly  $E[X_i] = P(X_i = 1)$ . Let us examine the probability  $P(X_i = 1)$ , i. e., that the officials agree on the  $i$ -th contestant. There are  $\binom{n}{2}$  pairs of officials. Let  $t_i$  denote the number of officials who say the  $i$ -th contestant is good and  $n - t_i$  be the number of those who do not. Then, there are  $\binom{t_i}{2}$  pairs who agree that the  $i$ -th contestant is good and  $\binom{n-t_i}{2}$  pairs who agree on the contrary. Altogether, there are  $\binom{t_i}{2} + \binom{n-t_i}{2}$  pairs that agree on the  $i$ -th contestant. Therefore,

$$E[X_i] = P(X_i = 1) = \frac{\binom{t_i}{2} + \binom{n-t_i}{2}}{\binom{n}{2}}.$$

Hence,

$$E[X] = \sum_{i=1}^m \frac{\binom{t_i}{2} + \binom{n-t_i}{2}}{\binom{n}{2}}.$$

We are going to show that for odd values of  $n$ , we have  $\binom{t_i}{2} + \binom{n-t_i}{2} \geq \frac{(n-1)^2}{4}$ . Rearranging this leads to

$$(n - 2t_i)^2 \geq 1 \Leftrightarrow t_i \leq \frac{n-1}{2} \text{ or } t_i \geq \frac{n+1}{2},$$

which is clearly true since  $\frac{n-1}{2}$  and  $\frac{n+1}{2}$  are adjacent integers.

Using the inequality  $\binom{t_i}{2} + \binom{n-t_i}{2} \geq \frac{(n-1)^2}{4}$ , we obtain:

$$E[X] \geq m \frac{\frac{(n-1)^2}{4}}{\frac{n(n-1)}{2}} = \frac{m(n-1)}{2n}.$$

Thanks to the assumption  $\frac{m(n-1)}{2n} > k$ , we have  $E[X] > k$ , and thus  $P(X > k) > 0$ , which finishes the proof.  $\square$

Further, we demonstrate application of probabilities to an interesting problem.

**13.G.3.** Let  $S$  be a finite set of points in the plane which are in general position (i. e., no three of them lie on a straight line). For any convex polygon  $P$  all of whose vertices lie in  $S$ , let  $a(P)$  denote the number of its vertices and  $b(P)$  the number of points from  $S$  which are outside  $P$ . Prove that for any real number  $x$ , we have

$$\sum_P x^{a(P)} (1-x)^{b(P)} = 1$$

where the sum runs over all convex polygons  $P$  with vertices in  $S$ . (A line segment, a singleton, and the empty set are considered to be a convex 2-gon, 1-gon, and 0-gon, respectively.)

KRUSKAL'S ALGORITHM

*Input:* A graph  $G = (V, E, w)$  with non-negative weights over edges.

*Output:* The minimal spanning trees for all components of  $G$ .

- (1) Sort the  $m$  edges in  $E$  so that  $w(e_1) \leq w(e_2) \leq \dots \leq w(e_m)$ .
- (2) For this order of the edges, call the "Spanning forest algorithm" from the previous subsection.

This is a typical example of the "greedy approach", when maximizing profits (or minimizing expenses) is achieved by choosing always the option which is the most advantageous at each stage.

In many problems, this approach fails since low expenses at the beginning may be the cause of much higher ones at the end. Therefore, greedy algorithms are often a base for very useful heuristic algorithms but seldom yield optimal solutions. However, in the case of minimum spanning tree, this approach works:

**Theorem.** *Kruskal's algorithm finds a minimum spanning tree for every connected graph  $G$  with non-negative edge weights. The algorithm runs in  $O(m \log m)$  time where  $m$  is the number of edges of  $G$ .*

**PROOF.** Let  $T = (V, E(T))$  denote the spanning tree generated by Kruskal's algorithm and, further, let  $\tilde{T} = (V, E(\tilde{T}))$  be an arbitrary minimum spanning tree. The minimality implies that  $\sum_{e \in E(\tilde{T})} w(e) \leq \sum_{e \in E(T)} w(e)$ , so the goal is to show that also  $\sum_{e \in E(T)} w(e) \leq \sum_{e \in E(\tilde{T})} w(e)$ .

If  $E(T) = E(\tilde{T})$ , then nothing further is needed. So assume there exists an edge  $e \in E(T)$  such that  $e \notin E(\tilde{T})$ . From all such edges, choose one, call it  $e$  with weight  $w(e)$  as small as possible.

The addition of  $e$  into  $\tilde{T}$  creates a cycle  $ee_1e_2 \dots e_k$  in  $\tilde{T}$ , and at least one of its edges  $e_i$  is not in  $E(T)$ . The choice of the edge  $e$  implies that if  $w(e_i) < w(e)$ , then the edge  $e_i$  would be among the candidate edges in Kruskal's algorithm after a certain subtree  $T' \subseteq T \cap \tilde{T}$  had been created, so its addition to the gradually constructed tree  $T$  would not create a cycle. Therefore, if  $w(e_i) < w(e)$ , the edge  $e_i$  would be chosen in the algorithm. It follows that  $w(e_i) \geq w(e)$ .

However, now the edge  $e_i$  can be replaced with  $e$  in  $\tilde{T}$  (by the choice of  $e_i$ , this results in a spanning tree again) without increasing the total weight. So the resulting  $\tilde{T}$  is a minimum spanning tree. It differs from  $T$  in fewer edges than before. Therefore, in a finite number of steps,  $\tilde{T}$  is changed to  $T$  without increasing the total weight.  $\square$

**13.2.8. Two more algorithms.** The second algorithm for finding a spanning tree, presented in 13.2.6 also leads to a minimum spanning tree:

was active in statistics, Clyde was a computer scientist too. The above algorithm dates from 1956.

**Solution.** First of all, we prove the wanted equality for  $x \in [0, 1]$ . Let us color a point from  $S$  so that it is white with probability  $x$  and black with probability  $1 - x$  (in other words, we consider a random choice of the size  $|S|$  with the binomial probability distribution  $\text{Bi}(n, x)$  and let us say that success corresponds to white and failure corresponds to black). We can note that for any such coloring, there must exist a polygon such that all of its vertices are white and all points outside are black (this polygon is the convex hull of the white points). The above suggests that the probability that the random choice realizes a polygon with all vertices white and all exterior points black is equal to one. However, we can compute this probability in a different way. The event of a polygon having this property is the union of  $k$  disjoint events, where  $k$  is the number of convex polygons, namely that a given polygon has the desired property (note that the property cannot be shared by different convex polygons). For every given convex polygon  $P$ , the probability that its vertices are white and the points outside it are black is equal to  $x^{a(P)}(1 - x)^{b(P)}$ , where  $a(P)$  is the number of vertices of  $P$  and  $b(P)$  is the number of points from  $S$  outside  $P$ . Since the probability of a union of disjoint events is equal to the sum of the particular events' probabilities, we get

$$\sum_P x^{a(P)}(1 - x)^{b(P)} = 1.$$

This proves the equality for all numbers in the interval  $[0, 1]$ . However, we can also perceive this fact as follows: any real number from the interval  $[0, 1]$  is a root of the polynomial  $\sum_P x^{a(P)}(1 - x)^{b(P)} - 1$ . As we know, a non-zero polynomial over the (infinite) field of real numbers can have only finitely many roots (see 12.3.6). Therefore,  $\sum_P x^{a(P)}(1 - x)^{b(P)} - 1$  is the zero polynomial and the equality  $\sum_P x^{a(P)}(1 - x)^{b(P)} = 1$  thus holds for all real numbers  $x$ .  $\square$

**Remark.** This equality holds even if we define the numbers  $a(P)$  and  $b(P)$  in another way: The definition of  $a(P)$  is the same, but now let  $b(P)$  denote the number of points from  $S$  which are not the vertices of  $P$ . (Thus, we always have  $a(P) + b(P) = |S|$ ). Then, the given equality is a corollary of the binomial theorem for  $(x + (1 - x))^{|S|}$ .

**13.G.4.** A competition with  $n$  players is called an  $(n, k)$ -tournament iff it has  $k$  rounds and satisfies the following:

- i) every player competes in every round and any pair of players competes at most once,
- ii) if  $A$  plays with  $B$  in the  $i$ -th round,  $C$  plays with  $D$  in the  $i$ -th round, and  $A$  plays with  $C$  in the  $j$ -th round, then  $B$  plays with  $D$  in the  $j$ -th round

Find all pairs  $(n, k)$  for which there exists an  $(n, k)$ -tournament.

**Solution.** There exists an  $(n, k)$ -tournament if and only if  $2^{\lceil \log_2(k+1) \rceil}$  divides the integer  $n$ . First of all, we are going to show the if-part. We construct a  $(2^t, k)$ -tournament

JARNÍK-PRIM'S ALGORITHM<sup>7</sup>

*Input:* A connected graph  $G = (V, E, w)$  with  $n$  vertices and  $m$  edges, with non-negative weights over the edges.

*Output:* The minimum spanning tree  $T$  of  $G$ .

- (1) Initialize  $T_0 = (\{v\}, \emptyset)$  with some vertex  $v \in V$ .
- (2) In the  $i$ -th step, look for all edges  $e$  which are not in  $T_{i-1}$ , but their tail vertex  $v_e$  is. Take the one of them with minimal weight and add it to  $T_{i-1}$ , i.e. add the head vertex to  $V_{i-1}$  and  $e$  to  $E_{i-1}$ .
- (3) The algorithm terminates when the number of added edges totals at  $n - 1$ .

The *Borůvka's algorithm* is similar. It constructs as many as possible connected components simultaneously: It begins with the singleton components in the graph  $T_0 = (V, \emptyset)$ . In each step, it connects every component to another component with the shortest edge possible. It is easy to prove that (provided the edge weights are pairwise distinct) this results in a minimum spanning tree.

BORŮVKA'S ALGORITHM

*Input:* A connected graph  $G = (V, E, w)$  with non-negative weights on the edges.

*Output:* The minimum spanning tree for  $G$ .

- (1) *Initialization.* Create the graph  $S$  with the same vertex set as  $G$  and no edges;
- (2) *The main loop.* While  $S$  contains more than one component, do:
  - for every tree  $T$  in  $S$ , find the shortest edge that connects  $T$  to  $G \setminus T$ , and add this edge into  $E$ ,
  - add all edges of  $E$  into the graph  $S$  and clear  $E$ .

Note that Borůvka's algorithm can be executed using parallel computation, which is why it is used in many practical modifications.

The proofs that both of these algorithms are correct, are similar to that of Kruskal's. The details are omitted.

13.2.9. Traveling salesman problem.



So far, our short excursion through graph based algorithms could give the feeling that simple and straightforward algorithms for the considered problems can always be found. So far however, only the easy problems have been considered. In all but very few cases, the contrary is true — mostly there are no algorithms running in polynomial time, so one needs to use algorithms which do not always find the optimal solution but give

<sup>7</sup>Robert Clay Prim (born 1921) is an American mathematician and computer scientist. While he published his work already in the realm of computer science and hence the most common name of the algorithm is "Prim's", earlier works by Otakar Borůvka (1899 - 1995) and Vojtěch Jarník (1897–1970) appeared before those by Prim. The Borůvka's algorithm was designed when consulting the construction of new electricity network in Moravia, a region in central Europe, in 1926, and Jarník published the algorithm (recovered much later by Prim) in 1930, motivated by Brůvka.

where  $k \leq 2^t - 1$  (then, the general case  $2^t \mid n$  can be easily derived from that). There are thus  $2^t$  players in the tournament, so we assign to each player a (unique) number from the set  $\{0, 1, \dots, 2^t - 1\}$ . In the  $i$ -th round, player  $\alpha$  competes with player  $\alpha \oplus i$  (where  $\oplus$  is the binary XOR operation, i. e., the  $j$ -th bit of  $a \oplus b$  is one if and only if the  $j$ -th bit of  $a$  is different from the  $j$ -th bit of  $b$ ). This schedule is correct since every player is engaged in every round and different players have different opponents (for  $\alpha \neq \beta$ , we have  $\alpha \oplus i \neq \beta \oplus i$ ). Further, the opponent of the opponent of  $\alpha$  is indeed  $\alpha$  (since  $(\alpha \oplus i) \oplus i = \alpha$ ). Moreover, the second tournament rule is also satisfied: if  $\alpha$  plays with  $\beta$  and  $\gamma$  plays with  $\delta$  in the  $i$ -th round (i. e.,  $\beta = \alpha \oplus i$  and  $\delta = \gamma \oplus i$ ) and if  $j$  is such that  $\alpha$  plays with  $\gamma$  in the  $j$ -th round, then we have  $\beta \oplus j = (\alpha \oplus i) \oplus j = (\alpha \oplus j) \oplus i = \gamma \oplus i = \delta$ , so  $\beta$  indeed plays with  $\delta$  in the  $j$ -th round. Any  $(2^t \cdot s, k)$ -tournament where  $s$  is odd can be obtained as  $s$  parallelized  $(2^t, k)$ -tournaments.

Now, we are going to show that the condition  $2^{\lceil \log_2(k+1) \rceil} \mid n$  is necessary as well. Consider the graph  $G_i$  whose vertices correspond to the players and edges are between the pairs who have played in or before the  $i$ -th round. Consider players  $A$  and  $B$  who play together in round  $i + 1$ . We want to show that we must have  $|T| = |\Delta|$  where  $T$  is the component of  $A$  and  $\Delta$  is the component of  $B$ . Actually, we show that any player of  $T$  competes with a player of  $\Delta$  in round  $i + 1$ . Thus, let  $C \in T$ , i. e., in  $G_i$ , there exists a path  $A = X_1, X_2, \dots, X_m = C$  such that  $X_j$  has played with  $X_{j+1}$ ,  $j = 1, \dots, m - 1$ , in or before the  $i$ -th round. Consider the sequence  $Y_1, Y_2, \dots, Y_m$ , where  $Y_k$  is the opponent of  $X_k$  in round  $i + 1$ ,  $k = 1, \dots, m$  (thus  $Y_1 = B$ ). Then, for any  $1 \leq j \leq m - 1$ , we have that  $X_j$  competes with  $Y_j$  and  $X_{j+1}$  competes with  $Y_{j+1}$  in round  $i + 1$  (by the definition of the sequence  $Y_1, \dots, Y_m$ ) and in a certain  $r$ -th round ( $1 \leq r \leq i$ ),  $X_j$  played with  $X_{j+1}$  (by the definition of the sequence  $X_1, \dots, X_m$ ). However, by the second tournament rule, this means that  $Y_j$  also played with  $Y_{j+1}$  in the  $r$ -th round, so the edge  $Y_j Y_{j+1}$  is contained in  $G_i$  for any  $1 \leq j \leq m - 1$ . Thus,  $Y_1, Y_2, \dots, Y_m$  is a path in  $G_i$ , so  $B = Y_1$  and  $Y_m$  lie in the same component ( $\Delta$ ). It can be deduced analogously that any player from  $\Delta$  competes with a player of  $T$  in round  $i + 1$ , and since every player plays exactly once in a given round, we must have  $|T| = |\Delta|$ . By the definition of a component, the component of  $A$  in  $G_{i+1}$  is equal to  $T \cup \Delta$ . Then, we have either  $T = \Delta$  (then, the component of  $A$  in  $G_{i+1}$  is  $T$ ), or  $T \cap \Delta = \emptyset$  (in this case, the component of  $A$  in  $G_{i+1}$  is the disjoint union  $T \cup \Delta$ ). Altogether, the component of  $A$  in  $G_{i+1}$  is either the same or twice as great as in  $G_i$ . Now, consider the components  $T_1, T_2, \dots, T_k$  of  $A$  in the respective graphs  $G_1, G_2, \dots, G_k$ . We have  $|T_1| = 2$  (since  $A$  had exactly one opponent in the first round) and for  $1 \leq i \leq k - 1$ , we have either  $|T_i| = |T_{i+1}|$ , or  $2|T_i| = |T_{i+1}|$ . Therefore, the number of vertices (players) of every component is a power of two, i. e.,  $|T_k| = 2^l$  for some  $l$ , and  $T_k \geq k + 1$  ( $A$  had different opponents in the  $k$  rounds). Hence,  $2^l \geq k + 1$ , i.

one which is as good as possible. This is called a heuristic approach.

One of the most important combinatorial problems of this class is the problem of finding a minimum Hamiltonian cycle. This is a Hamiltonian cycle with the minimum sum of the weights of its edges among all Hamiltonian cycles.

This problem arises in many practical applications. For instance:

- goods or post delivery (via a given network)
- network maintenance (electricity, water pipelines, IT, etc.)
- request processing (parallel requests for reading from a hard disk, for instance),
- measuring several parts of a system (for example, when studying the structure of a protein crystal using X-rays, the main expenses are due to the movements and focusing for particular measurements),
- material division (for instance, when covering a wall with wallpaper, one tries to keep the pattern continuous while minimizing the amount of unused material)).

The greedy approach can be applied in case of looking for a minimum Hamiltonian cycle as well. The algorithm begins in an arbitrary vertex  $v_1$ , which is set active, and the other vertices are labeled as sleeping. For each step, it examines the sleeping vertices adjacent to the active one and selects the one which is connected by the shortest edge. The active vertex is labeled as processed, and the selected vertex becomes active. The algorithm terminates either with a failure, when there is no edge going from the active vertex to a sleeping one, or it successfully finds a Hamiltonian path. In the latter case, if there exists an edge from the last vertex  $v_n$  to  $v_1$ , a Hamiltonian cycle is obtained.

This algorithm seldom produces a minimal Hamiltonian cycle. At least, it always finds some (and relatively small) Hamiltonian cycle in a complete graph.

**13.2.10. Flow networks.** Another group of applications of the language of graph theory concerns moving some amount of a measurable material in a fixed network. The vertices of a directed graph represent places between which one transports material up to predetermined limits which are given as assessments of the edges (called capacities). There are two important types of vertices: *the sources and sinks of the network*. A network is a directed graph with valued edges, where some of the vertices are labeled as sources or sinks.

Without loss of generality, assume that the graph is directed and has only one source and one sink: In the general case, an artificial source and a sink can always be added, connected with directed edges to the original sources and sinks. Then the capacities of the added edges would cover all maximum capacities of the particular sources and sinks. The situation is depicted in the diagram. There, the black vertices on the left correspond to the given sources, while the black vertices on the right stand for the given sinks. On the left,



e.,  $2^l$  is at least  $2^{\lceil \log_2(k+1) \rceil}$ , so the number of players in each component is divisible by  $2^{\lceil \log_2(k+1) \rceil}$ . Thus, so must be the total number  $n$ .  $\square$

**H. Combinatorial games**

**13.H.1.** Consider the following game for two players: On the table, there are four piles of 9, 10, 11, and 14 tokens, respectively. Players alternate moves where the move consists of selecting one of the piles and removing an arbitrary (positive) number of tokens from that pile. The player who takes the last token wins. Is there a winning strategy for one of the players?

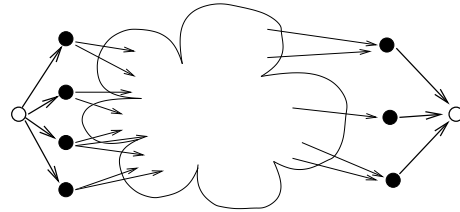
**Solution.** Note that this game is the sum of four games which correspond to one-pile games where an arbitrary (positive) number of tokens can be removed (the sum of combinatorial games is both commutative and associative, so we can talk just about the sum of those games without having to specify the order). A simple induction argument shows that the value of the Sprague-Grundy function (the *SG-value*) of such one-pile game is equal to the number of tokens: Suppose that a natural number  $n$  is such that for all  $k < n$ , the *SG-value* of the game with  $k$  tokens is  $k$ . According to the rules of the game, we can remove an arbitrary (positive) number of tokens, i. e., we can leave there an arbitrary number from 0 to  $n - 1$ . By the induction hypothesis, this means that for any number  $k < n$ , we can reach a position whose *SG-value* is  $k$ , and we cannot reach a position whose *SG-value* would be  $n$ . By the definition of the *SG-function*, the value of the game with  $n$  tokens is  $n$ . It follows from the theorem of subsection 13.2.16 that the *SG-value* of the initial position of our game is equal to the xor of the initial positions of the particular games, namely

$$9 \oplus 10 \oplus 11 \oplus 14 = 6.$$

Since this value is non-zero, there exists a winning strategy for the first player: he always moves to a position whose *SG-value* is zero—such a position must exist by the definition of the *SG-function*. For instance, the first move would be to remove 6 tokens from the pile containing 14. (We look at the highest one in the binary expansion of the *SG-value* and find a pile where the corresponding bit is also one. Then, we set this bit to zero—thereby surely decreasing the number of tokens—and adjust the lower bits so that there would be an even number of ones in each position, resulting in zero *SG-value*.)  $\square$

**13.H.2.** Consider the following game for two players: On the table, there is a pile of tokens. Players alternate moves where the move consists of either splitting one pile into two (non-empty) piles or removing an arbitrary (positive) number of tokens from a pile. The player who takes the last token wins. Find the *SG-value* of the initial position of this game if the pile contains  $n$  tokens.

there is an artificial source (a white vertex), and there is an artificial sink on the right. The edge values are not shown in the diagram.



FLOW NETWORKS

A network is a directed graph  $G = (V, E)$  with a distinctive vertex  $z$ , called the *source*, and another distinctive vertex  $s$ , called the *sink*, together with a non-negative assessment of the edges  $w : E \rightarrow \mathbb{R}$ , which represents their *capacities*. A *flow* in a network  $S = (V, E, z, s, w)$  is an assessment of the edges  $f : E \rightarrow \mathbb{R}$  such that, for each vertex  $v$  except for the source and the sink, the total input is equal to the total output, i.e.

$$\sum_{e \in IN(v)} f(e) = \sum_{e \in OUT(v)} f(e).$$

This rule is often called the Kirchhoff's law (referring to the terminology used in physics).

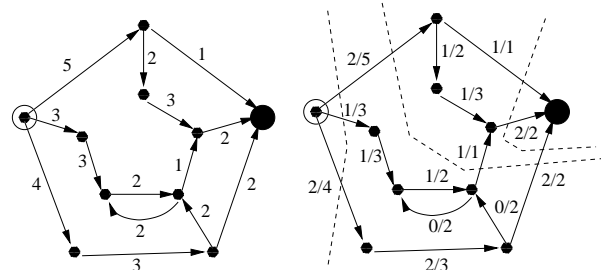
The *size of a flow*  $f$  is given by the total balance of the source values

$$|f| = \sum_{e \in OUT(z)} f(e) - \sum_{e \in IN(z)} f(e).$$

It follows directly from the definition that the size of a flow  $f$  can also be computed as

$$|f| = \sum_{e \in IN(s)} f(e) - \sum_{e \in OUT(s)} f(e).$$

The left hand part of the following diagram shows a simple network with the source in the white circled vertex and the sink in the black bold vertex. The labels over the edges determine the maximal capacities. Looking at the sum of the capacities that enter the sink, the maximum flow in this network is 5 (the sum of the capacities leaving the source is larger).



**13.2.11. Maximum flow problem.** The next task is to find the maximum possible flow for a given network on a graph  $G$ . The right hand side of the above diagram shows a flow of size five, and the size of any flow cannot exceed this. The fundamental principle is that the capacities of a set of edges

**Solution.** We are going to prove by induction that any positive integer  $k$  satisfies:

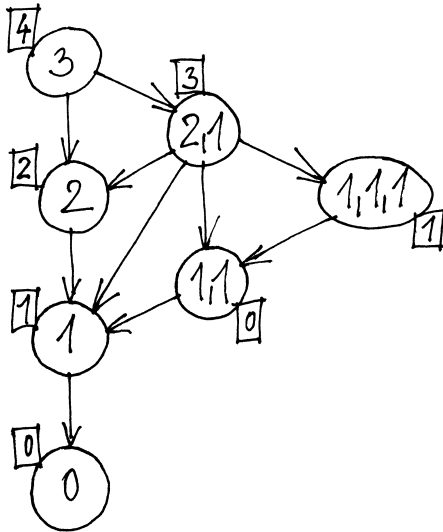
$$g(4k + 1) = 4k + 1$$

$$g(4k + 2) = 4k + 2$$

$$g(4k + 3) = 4k + 4$$

$$g(4k + 4) = 4k + 3$$

Clearly, we have  $g(0) = 0$ . The following picture shows how we can deduce the value of the  $SG$ -function for one-, two-, and three-token piles. However, it is apparent that this would be much harder for a general number of tokens.



Now, assume that the above is satisfied for all positive integers below  $4k + 1$  and let us prove that we indeed have  $g(4k + 1) = 4k + 1$ . By the definition, the  $SG$ -value is the least non-negative integer  $l$  such that there is no move to a position with  $SG$ -value  $l$ . Moreover, this property (including that the terminal positions have zero value) determine the Sprague-Grundy function uniquely. Therefore, it suffices to prove that, for each  $l < 4k + 1$ , we can move to a position with  $SG$ -value  $l$ , and that we cannot get into a position with  $SG$ -value  $4k + 1$ . The former is clear since by the induction hypothesis, the  $SG$ -values of one-pile games of  $0, 1, \dots, 4k$  tokens take all the integers  $0, 1, \dots, 4k$  (although not in this order), so we can just remove the corresponding number of tokens from the pile. Now, we are going to show that we cannot reach a position with  $SG$ -value  $4k + 1$ : We already know that the only moves that could possibly lead to this  $SG$ -value are to split the pile into two. If we examine the resulting amounts modulo 4, there are two possibilities: either the number of tokens in one of the resulting piles is divisible by 4 ( $4a$ ) and the other one leaves remainder 1 ( $4b+1$ ), or the numbers leave remainders 2 and 3, respectively. As for the former case, the  $SG$ -values of the resulting piles are, by the induction hypothesis,  $4a - 1$  and  $4b + 1$  (the numbers of tokens in the particular piles are non-zero and less than  $4k+1$ , so we may use the induction hypothesis. In the latter case, i.

through which each path from  $z$  to  $s$  must go have to be added up. In the diagram, there are three such choices providing the limits 12, 8, 5 (from left to right). At the same time, in such a simple case the flow that realizes the maximal possible value is easily found. This idea can be formalized as follows:

CUT IN A NETWORK

A *cut in a network*  $S = (V, E, z, s, w)$  is a minimal set of edges  $C \subseteq E$  such that when these edges are removed, there remains no path from the source  $z$  to the sink  $s$  in the graph  $G = (V, E \setminus C)$ . The number

$$|C| = \sum_{e \in C} w(e)$$

is called the *capacity of the cut*  $C$ .

Clearly, there is no flow whose size is greater than the capacity of a cut. We present the *Ford-Fulkerson algorithm*<sup>8</sup>, which finds a cut with the minimum possible capacity as well as a flow which realizes this value. This proves the following theorem:

**Theorem.** In any network  $S = (V, E, z, s, w)$ , the maximum size of a flow equals the minimum capacity of a cut in  $S$ .

The idea of the algorithm is quite simple. It looks for paths between the vertices of the graph, trying to “saturate” them with the flow. For this purpose, define the following terminology:

An undirected path from the vertex  $v$  to the vertex  $v'$  in a network  $S = (V, E, z, s, w)$  is called *unsaturated* if and only if all edges  $e$  directed along the path from  $v$  to  $v'$  satisfy  $f(e) < w(e)$  and the edges  $e$  in the other direction satisfy  $f(e) > 0$  (sometimes, one tries to saturate the flow in the other direction; yielding a *semipath*, or the *augmenting semipath*). The *residual capacity* of an edge  $e$  is the number  $w(e) - f(e)$  if the edge is directed from  $v$  to  $w$ , and it is the number  $f(e)$  otherwise. The residual capacity of a path is defined to be the minimum residual capacity of its edges. For the sake of simplicity, assume that all the edge capacities are rational numbers.

<sup>8</sup>Ford, L. R.; Fulkerson, D. R. (1956). "Maximal flow through a network". Canadian Journal of Mathematics 8: 399–404.



e., if we split the pile into  $4a+2$  and  $4b+3$  tokens, we get that their  $SG$ -values are  $4a + 2$  and  $4b + 4$ . Furthermore, a two-pile game is the sum of the two corresponding one pile-games, so the  $SG$ -value of the two-pile game is the xor (nim-sum) of the amounts. In both cases, the  $SG$ -value leaves remainder 2 upon division by 4 (consider the last two bits). In particular, it is surely not equal to  $4k + 1$ . This proves the induction step for positive integers of the form  $4k + 1$ .

The proof for integers of the form  $4k + 2$  is analogous. The situation is more amazing in the  $4k + 3$  case: Similarly as above, it follows from the induction hypothesis that the  $SG$ -values of the one-pile positions we can move to exhaust all the non-negative integers up to  $4k + 2$ . However, note that if we split the pile into two containing 1 and  $4k + 2$  tokens, respectively, then their  $SG$ -values are also 1 and  $4k+2$  by the induction hypothesis, and the xor of these integers is  $4k + 3$ . It remains to prove that there is no move into a position with  $SG$ -value  $4k + 4$ : Again, the only remaining possibility is to split the existing pile. Then, the resulting remainders modulo 4 are either 0 and 3, or 1 and 2. By the induction hypothesis the remainders of the corresponding  $SG$ -values are respectively 3 and 0 in the former case, and 1 and 2 in the latter. In either case, the xor of these integers (and thus the  $SG$ -value of the resulting position) leaves remainder 3, so it is not equal to  $4k + 4$ . This proves the induction step for positive integers of the form  $4k + 3$ . The proof for integers of the form  $4k + 4$  is analogous.  $\square$

**I. Generating functions**

**13.I.1.** In how many ways can we buy 12 packs of coffee if we can choose from 5 kinds?

Further, solve this problem with the following modifications:

- i) we want to buy at least 2 packs of each kind;
- ii) we want to buy an even number of packs of each kind;
- iii) there are only 3 packs of one of the kinds.

**Solution.** The basic problem is a classical example of a combinatorial problem on the number of 5-combinations with repetition—the answer is  $\binom{12+5-1}{5-1} = \binom{16}{4}$ . The modifications can also be solved by combinatorial reasonings with a bit of invention. However, we want to demonstrate how these problems can be solved (almost without no need to think) using generating functions.

The wanted number corresponds to the coefficient at  $x^{12}$  in the expansion of the function

$$(1 + x + x^2 + \dots)^5 = (1 + x + \dots)(1 + x + \dots) \dots (1 + x + \dots)$$

into a power series. The number of packs of the first kind determines which term is selected from the first parenthesis, and similarly for the other kinds. (Note that we need not pay special attention to that fact that there cannot be more than 12 packs of a given kind – it turns out that infinite series are usually simpler to work with than finite polynomials.)

**FORD-FULKERSON ALGORITHM**

*Input:* A network  $S = (V, E, z, s, w)$ .

*Output:* A maximal possible flow  $f : E \rightarrow \mathbb{R}$  and a minimal cut  $C$ , which is given by those edges which lead from the set  $U \subseteq V$  of those vertices to which there exists an unsaturated path, to  $V \setminus U$ .

- (1) *Initialization:* Set  $f(e) = 0$  for each edge  $e \in E$ , and using depth-first search from  $z$ , find the set  $U \subseteq V$  of those vertices to which there exists an unsaturated path.
- (2) *The main loop:* While  $s \in U$ , do
  - select an unsaturated path  $P$  from the source  $z$  to the sink  $s$ ; then increase the flow  $f$  along all edges of the path  $P$  by the value of the residual capacity of  $P$ .
  - update  $U$ .

**PROOF.** As seen, the size of any flow cannot exceed the capacity of any cut. Therefore, it suffices to show that when the algorithm terminates, the capacity of the generated cut equals the size of the constructed flow.

The algorithm terminates in the first moment when there is no unsaturated path from the source  $z$  to the sink  $s$ . This means that  $U$  does not contain  $s$  and for all edges  $e$  from  $U$  and ending outside of  $U$ ,  $f(e) = w(e)$  (otherwise, the other endpoint of  $e$  would be added to  $U$  when updating  $U$ ).

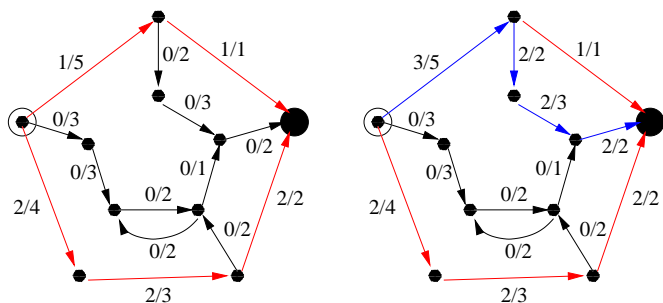
For the same reason, all edges  $e$  leading from  $V \setminus U$  to  $U$  must have  $f(e) = 0$ .

Clearly, at each beginning of the main loop the total size of the flow satisfies

$$|f| = \sum_{\text{edges from } U \text{ to } V \setminus U} f(e) - \sum_{\text{edges from } V \setminus U \text{ to } U} f(e)$$

However, when the algorithm terminates, this expression equals to the capacity of the final cut  $C$  which is the desired result.

It remains to show that the algorithm always terminates. Since the edges are assumed assessed with rational numbers, it can be assumed by rescaling that the capacities are integers. Then every flow constructed during the run of the algorithm has integer size. In addition, every iteration of the main loop increases the size of the flow. However, since any cut bounds the maximum size of any flow from above, the algorithm must terminate after a finite number of steps.  $\square$



Since

$$\frac{1}{1-x} = 1 + x + x^2 + \dots$$

(see 13.4.3), the function we are considering is  $(1-x)^{-5}$ . Our task is thus to expand  $(1-x)^{-5}$  into a power series. By the generalized binomial theorem, from 13.4.3, the coefficient at  $x^k$  is the number  $\binom{k+5-1}{5-1}$ , which is  $\binom{16}{4}$  in our case. Note that using generating functions, we have answered the question not only for 12, but rather for an arbitrary number of packs of coffee.

The modifications can be solved analogously:

- i) The generating function is

$$(x^2 + x^3 + \dots)^5 = \left(\frac{x^2}{1-x}\right)^5 = \frac{x^{10}}{(1-x)^5},$$

hence the coefficient at  $x^{12}$  is equal to  $\binom{2+5-1}{5-1}$ .

- ii) An even number of each kind corresponds to the generating function

$$(1 + x^2 + x^4 + \dots)^5 = \frac{1}{(1-x^2)^5}.$$

The coefficient at  $x^{12}$  can be found by many means; the easiest one seems to be the substitution  $y = x^2$  and looking for the coefficient at  $y^6$  (which can be perceived as joining the packs into pairs in the shop). This leads to the answer  $\binom{6+5-1}{5-1}$ .

- iii) In this case, the generating function equals

$$(1 + x + x^2 + x^3)(1 + x + x^2 + \dots)^4,$$

and the wanted result is thus

$$\binom{12+4-1}{4-1} + \binom{11+4-1}{4-1} + \binom{10+4-1}{4-1} + \binom{9+4-1}{4-1}.$$

**13.I.2.** In how many ways can we use the coins of values 1, 2, 5, 10, 20, and 50 crowns to pay exactly 100 crowns?

**Solution.** We are looking for non-negative integers  $a_1, a_2, a_5, a_{10}, a_{20}$ , and  $a_{50}$  such that  $a_i$  is a multiple of  $i$  for all  $i \in \{1, 2, 5, 10, 20, 50\}$  and, at the same time,  $a_1 + a_2 + a_5 + a_{10} + a_{20} + a_{50} = 100$ . We can see that the wanted number of ways can be obtained as the coefficient at  $x^{100}$  in the product

$$\begin{aligned} &(1 + x + x^2 + \dots)(1 + x^2 + x^4 + \dots)(1 + x^5 + x^{10} + \dots) \\ &\cdot (1 + x^{10} + x^{20} + \dots)(1 + x^{20} + x^{40} + \dots) \\ &\cdot (1 + x^{50} + x^{100} + \dots) = \\ &= \frac{1}{1-x} \cdot \frac{1}{1-x^2} \cdot \frac{1}{1-x^5} \cdot \frac{1}{1-x^{10}} \cdot \frac{1}{1-x^{20}} \cdot \frac{1}{1-x^{50}}. \end{aligned}$$

The result can be obtained using the software of SAGE, for instance (the names of the used commands are pretty self-descriptive, aren't they?):

```
sage : f=1/(1-x)*1/(1-x^2)*1/(1-x^5)\
      *1/(1-x^10)*1/(1-x^20)*1/(1-x^50)
sage : r=taylor(f,x,0,100)
sage : r.coeff(x,100)

4562
```

The run of the algorithm is illustrated in two diagrams. On the left, there are two shortest unsaturated paths from the source to the sink in gray (the upper one has two edges, while the lower one has three). On the right, another path is saturated (taking the first turn in the upper path), also drawn in gray. Now, it is apparent that there can be no other unsaturated path from the source to the sink. Therefore, the algorithm terminates at this moment.

**13.2.12. Further remarks.** The algorithm allows for further conditions incorporated in the problem. For instance, capacity limits can be set for the vertices of the network as well. There are not only upper limits for the flows along particular edges or through vertices, but also lower ones.



It is easy to add vertex capacities – just double every vertex (one for the incoming edges, the other for the outgoing edges), connecting each pair with an edge of the corresponding capacity.

The lower limits for the flow can be included in the initialization part of our algorithm. However, one needs to check whether such a flow exists at all. Many other variations can be found in literature.

On the other hand, the algorithm does not necessarily terminate if the edge capacities are irrational. Moreover, the flows that are constructed during the run may not even converge to the optimal solution in such a case. However, it still holds that if the algorithm terminates, then a maximum flow is found.

If the capacities are integers (equivalently rational numbers), the running time of the algorithm can be bounded by  $O(f|E|)$ , where  $f$  is the size of a maximum flow in the network and  $|E|$  is the number of edges. The worst case occurs if every iteration increases the size of the flow by one.

In the proof of correctness, no explicit way of searching the graph when looking for an unsaturated path is used. Another variation of the Ford–Fulkerson algorithm is to use breadth-first search. The resulting algorithm is called Edmonds–Karp, and its running time is  $O(|V|^2|E|)$ .<sup>9</sup> We mention Dinic's algorithm, which simplifies the search for an unsaturated path by constructing the *level graph*, where *augmenting edges* are considered only if they lead between vertices whose distances from the source differ. The time complexity of this algorithm is  $O(|V|^2|E|)$ , which is much better for dense graphs than the Edmonds–Karp algorithm.

**13.2.13. Problems related to flow networks.** A good application of flow networks is the problem of *bipartite matching*. The task is to find a maximum matching in a bipartite graph, i.e. a set of as many edges as possible so that each vertex of the graph is the endpoint of at most one of the selected edges.

<sup>9</sup>Edmonds, Jack; Karp, Richard M. (1972). "Theoretical improvements in algorithmic efficiency for network flow problems". *Journal of the ACM (Association for Computing Machinery)* 19 (2): 248–264. doi:10.1145/321694.321699.

Put some nice example in the other column, e.g. the <https://www.cs.princeton.edu/courses/archive/spring13/cos423/lectures/07DemoFordFulkersonPathological.pdf>

**13.I.3.** Expand the following functions into power series:

- i)  $\frac{x}{x+2}$ ,
- ii)  $\frac{x^2+x+1}{2x^3+3x^2+1}$ .

**Solution.**

i)

$$\begin{aligned} \frac{x}{x+2} &= \frac{x}{2 - (-x)} = \frac{x/2}{1 - (-x/2)} = \\ &= \frac{x}{2} - \frac{x^2}{4} + \frac{x^3}{8} - \dots + \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{2^n}. \end{aligned}$$

ii) We perform partial fraction decomposition:

$$\begin{aligned} \frac{x^2+x+1}{2x^3+3x^2+1} &= \frac{x^2+x+1}{(x-1)^2(2x+1)} = \\ &= \frac{A}{2x+1} + \frac{B}{x-1} + \frac{C}{(x-1)^2}, \end{aligned}$$

finding out that  $A = B = \frac{1}{3}$  and  $C = 1$ ; hence

$$\begin{aligned} \frac{x^2+x+1}{2x^3+3x^2+1} &= \frac{1/3}{1+2x} - \frac{1/3}{1-x} + \frac{1}{(1-x)^2} = \\ &= \sum_{n=0}^{\infty} \left[ \frac{1}{3} ((-2)^n - 1) + (n+1) \right] x^n \end{aligned}$$

**13.I.4.** Find the generating functions of the following sequences:

- i)  $(1, 2, 3, 4, 5, \dots)$ ,
- ii)  $(1, 4, 9, 16, \dots)$ ,
- iii)  $(1, 1, 2, 2, 4, 4, 8, 8, \dots)$ ,
- iv)  $(9, 0, 0, 2 \cdot 16, 0, 0, 4 \cdot 25, 0, 0, 8 \cdot 36, \dots)$ ,
- v)  $(9, 1, -9, 32, 1, -32, 100, 1, -100, \dots)$ .

**13.I.5.** In how many ways can we buy  $n$  pieces of the following five kinds of fruit if we do not distinguish between particular pieces of a given kind, we need not buy all kinds, and:

- there is no restriction on the number of apples we buy,
- we want to buy an even number of bananas,
- the number of pears we buy must be a multiple of 4,
- we can buy at most 3 oranges, and
- we can buy at most 1 pomelo.

- This is an abstract variation of a quite common problem. For instance, it may be needed to match boys and girls in dancing lessons, provided information about which pairs would be willing to dance together is given.

This problem is easily reduced to the problem of maximum flow. Add an artificial source to the graph and connect it with edges to all vertices of one part of the bipartite graph, while the vertices of the other part are connected to an artificial sink. The capacity of each edge is set to one, and the resulting graph is searched for the maximum flow. Then, the edges that are used in the flow correspond to the selected pairs. Of course, information on which pairs to put together by leaving some of them out, may be included.

Another important application of flow networks is the proof of Menger's theorem (mentioned as a theorem in 13.1.10). It can be understood as follows: Given a directed graph, set the capacity of each edge as well as edge vertex to one. Further, select an arbitrary pair of vertices  $v$  and  $w$ , which are considered to be the source and the sink, respectively. Then, the size of a maximum flow in this graph equals the maximum number of disjoint paths from  $v$  to  $w$  (the paths may share only the source and the sink). Every cut divides  $v$  and  $w$  into different connected components of the remaining graph (since they are chosen to be the source and sink). The desired statements then follow from the fact that the size of a maximum flow equals the capacity of a minimum cut.



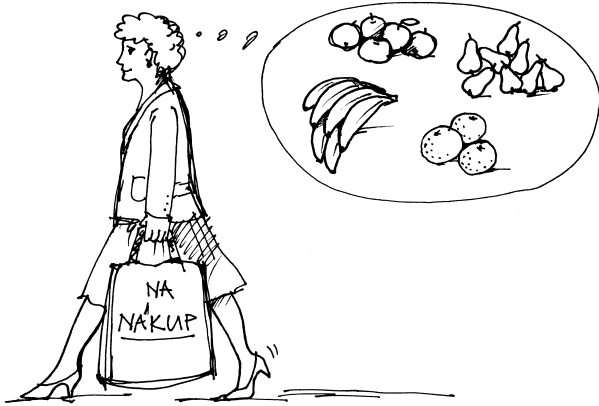
**13.2.14. Game trees.** We turn our attention to a very broadly used application of tree structures when analyzing possible strategies or procedures. They can be encountered in the theory of artificial intelligence as well as in the game theory. They play an important role in economics and many other social fields.

This is about *games*. In the mathematical sense, game theory examines models in which one or more players take turns in playing moves according to predetermined and generally known rules. Usually, the moves are assessed with profits or losses for the given player. Then, the task is to find a *strategy* for each player, i.e. an algorithmic procedure which maximizes the profits or minimizes the losses.

We use an extensive description of the games. This means that a complete and finite analysis of all possible states of the game is given, and the resulting analysis gives an exact account about the profits and losses. This is supposing that the other players also play the best moves for them.

A *game tree* is a rooted tree whose vertices are the possible states of the game and they are labeled according to whose turn it is. The outgoing edges of a vertex correspond to the possible moves of the player from that state. This complete description of a game using the game tree may be used for common games like chess, naughts and crosses (known also as tic-tac-toe), etc.





**Solution.** The generating function for the sequence  $(a_n)$  where  $a_n$  is the wanted number of ways to buy  $n$  pieces of fruit is

$$\begin{aligned} & (1 + x + x^2 + \dots)(1 + x^2 + x^4 \dots)(1 + x^4 + x^8 + \dots) \cdot \\ & \cdot (1 + x + x^2 + x^3)(1 + x) = \\ & = \frac{1}{1-x} \cdot \frac{1}{1-x^2} \cdot \frac{1}{1-x^4} \cdot \frac{1-x^4}{1-x} \cdot (1+x) = \\ & = \frac{1}{(1-x)^3}. \end{aligned}$$

By the generalized binomial theorem, we have  $(1-x)^{-3} = \sum_{n=0}^{\infty} \binom{n+2}{2} x^n$ . Therefore, the wanted number of ways satisfies  $a_n = \binom{n+2}{2}$ .  $\square$

**13.I.6.** Using the generalized binomial theorem, prove again the following combinatorial identities:

- $\sum_{k=0}^n \binom{n}{k} = 2^n$ ,
- $\sum_{k=0}^n (-1)^k \binom{n}{k} = 0$ ,
- $\sum_{k=0}^n k \binom{n}{k} = n2^{n-1}$ .

**Solution.** Substituting into the binomial theorem

$$(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \dots + \binom{n}{n}x^n$$

the numbers  $x = 1$  and  $x = -1$ , we obtain the first and second identities, respectively. Then, the third one can be obtained by viewing both sides of the binomial theorem “continuously” and using the properties of derivatives.  $\square$

**13.I.7.** In a box, there are 30 red, 40 blue, and 50 white balls. Balls of one color are indistinguishable. In how many ways can we select 70 balls?

**Solution.** The wanted number is equal to the coefficient at  $x^{70}$  in the product

$$(1 + x + \dots + x^{30})(1 + x + \dots + x^{40})(1 + x + \dots + x^{50}).$$

This product can be rearranged to

$$(1-x)^{-3}(1-x^{31})(1-x^{41})(1-x^{51}),$$

whence, using the generalized binomial theorem, we obtain

$$\left( \binom{2}{2} + \binom{3}{2}x + \binom{4}{2}x^2 + \dots \right) (1-x^{31}-x^{41}-x^{51}+x^{72}+\dots).$$

As a simple example, consider a simple variation of the game known as *Nim*.<sup>10</sup>

There are  $k$  tokens on the table (the tokens may be sticks or matches), where  $k > 1$  is an integer, and players take turns at removing one or two tokens. The player who manages to take the last token(s) wins. There is a variation of the game in which the player who is forced to take the last token loses. The tree of this game, including all necessary information about the game, can be constructed as follows:

- The state with  $\ell$  tokens on the table and the first player to move corresponds to the subtree rooted at  $F_\ell$ . The state with the same number of tokens but the second player to move is represented by the subtree rooted at  $S_\ell$ .
- The vertex  $F_\ell$  has  $S_{\ell-1}$  as its left-hand son and  $S_{\ell-2}$  as its right-hand son. Similarly, the sons of the vertex  $S_\ell$  are  $F_{\ell-1}$  and  $F_{\ell-2}$ .
- The leaves are always  $F_0$  or  $S_0$ . (In the variation when the player to take the last token loses, these would be the states  $F_1$  and  $S_1$ .)

Every run of the game starting at root  $F_k$  corresponds to exactly one leaf of the resulting tree. Therefore, the total number  $p(k)$  of possible runs for  $F_k$  is equal to

$$p(k) = p(k-1) + p(k-2)$$

for  $k \geq 3$ , and clearly  $p(1) = 1$  and  $p(2) = 2$ . This difference equation is already considered. It is satisfied by the Fibonacci numbers, which can be computed by an explicit formula (see the subsection on generating functions in the end of this chapter, or the corresponding part about difference equations in chapter three, cf. 3.B.1). A formula is known for the number of possible runs of the game. The number of possible states equals the number of all vertices of the tree. The game always ends in a win of one of the players. We can also consider games where a tie is possible.

**13.2.15. Game analysis.** The tree structure allows an analysis of the game so that an algorithmic strategy for each player can be built. This is done with a simple recursive procedure for assessing the root of a subtree. Each vertex is given a label:  $W$  for vertices where the first player can force a win,  $L$  for those where the first player loses if the other one plays optimally, and, optionally,  $T$  for vertices where optimal play of both players results in a tie. The procedure is as follows:

- (1) The leaves are labeled directly according to the rules of the game (in the case of our Nim, the leaves  $S_0$  are labeled by  $W$ , and the leaves  $F_0$  by  $L$ ).
- (2) Considering the vertex  $F_\ell$ . Label it  $W$  if it has a son who is labeled by  $W$ . If there is no such son, but there is a son labeled by  $T$ , then  $F_\ell$  is given the label  $T$ . Otherwise, i.e. if all sons are labeled by  $L$ , then  $F_\ell$  also gets  $L$ .



<sup>10</sup>The game was given this name by Charles Bouton in his analysis of this type of games from 1901. It refers to the German word “Nimm!”, meaning “Take!”.

Hence, the coefficient at  $x^{70}$  is clearly  $\binom{70+2}{2} - \binom{70+2-31}{2} - \binom{70+2-41}{2} - \binom{70+2-51}{2} = 1061$ .  $\square$

**13.I.8.** Prove that

$$\sum_{k=1}^n H_k = (n+1)(H_{n+1} - 1).$$

**Solution.** The necessary convolution can be obtained as the product of the series  $\frac{1}{1-x}$  and  $\frac{1}{1-x} \ln \frac{1}{1-x}$ . Hence:

$$[x^n] \frac{1}{(1-x)^2} \ln \frac{1}{1-x} = \sum_{k=1}^n \frac{1}{k} (n+1-k),$$

whence the wanted identity follows easily.  $\square$

**13.I.9.** Solve the recurrence

$$\begin{aligned} a_0 &= a_1 = 1, \\ a_n &= a_{n-1} + 2a_{n-2} + (-1)^n. \end{aligned}$$

**Solution.** As always, it may be a good idea to write out a few terms of the sequence (however, this will not help us much in this case; still, it can serve as verification of the result).<sup>2</sup>

Step 1:  $a_n = a_{n-1} + 2a_{n-2} + (-1)^n [n \geq 0] + [n = 1]$ .

Step 2:  $A(x) = xA(x) + 2x^2A(x) + \frac{1}{1+x} + x$ .

Step 3:

$$A(x) = \frac{1+x+x^2}{(1-2x)(1+x)^2}.$$

Step 4:  $a_n = \frac{7}{9}2^n + (\frac{1}{3}n + \frac{2}{9})(-1)^n$ .  $\square$

**13.I.10. Quicksort – analysis of the average case.** Our task is to determine the expected number of comparisons made by Quicksort, a well-known algorithm for sorting a (finite) sequence of elements.

An example of a simple *divide-and-conquer* implementation:

```

if L == []: return []
return qsort ([ x for x in L[1:] if x < L
[0]])
+ L[0:1]
+ qsort ([ x for x in L[1:] if x >= L
[0]])
    
```

It is not too difficult to construct a formula for the number of comparisons (we assume that the particular orders of the sequence to be sorted are distributed uniformly). The following parameters are important for the analysis of the algorithm:

- i) The number of comparisons in the *divide* phase:  $n - 1$ .
- ii) The uniformness assumption: the probability that  $L[0]$  is the  $k$ -th greatest element of the sequence is  $\frac{1}{n}$ .
- iii) The sizes of the sorted subsequences in the *conquer* phase:  $k - 1$  and  $n - k$ .

<sup>2</sup>Despite the statement in *Concrete mathematics*, this sequence can already be found in *The On-Line Encyclopedia of Integer Sequences*.

(3) Similarly, a vertex  $S_\ell$  is labeled  $L$  if it has a son labeled by  $L$ . Otherwise if it has a son labeled by  $T$ , it receives  $T$ . Otherwise (i.e. if it has only  $W$ -sons), it is labeled by  $W$ .

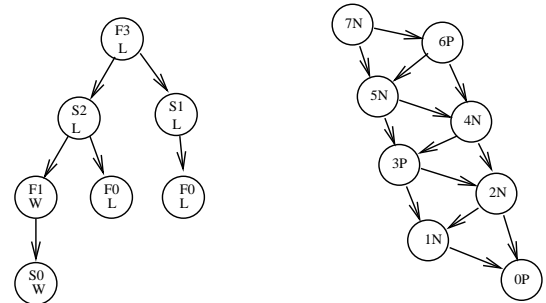
Calling this procedure on the root of the tree gives the labeling of each vertex as well as an optimal strategy for each player:

- The first player tries to move to a vertex labeled by  $W$ ; if this cannot be achieved, he moves to a  $T$ -vertex at least.
- Similarly, the second player tries to move to a vertex labeled by  $L$ ; if this cannot be achieved, he moves to a  $T$ -vertex at least.

The depth of the recursion is given by the depth of the tree. For instance, having a Nim game with  $k$  tokens, the depth is  $k$ .

This analysis is not very useful yet. In order to use it in the mentioned form, the entire game tree is needed for disposal. This can be a great amount of data (for instance, in the case of naughts and crosses on  $3 \times 3$  playground, the corresponding tree has tens of thousands of vertices). Usually, the analysis with game tree is used when only a small part of the whole tree is examined, applying appropriate heuristic methods, and the corresponding part is being created dynamically during the game. This is a fascinating field of the modern theory of artificial intelligence. The details are omitted.

There is a more compact representation of the tree structure for our purposes of complete formal analysis. If the game tree for Nim is drawn, then one state of the game is represented by many vertices which correspond to different histories of the game. However, the strategies depend only on the actual state (i.e. the number of tokens and the player to move) rather than on the history of the game. Therefore, the same game can be described by a graph where for each number of tokens, there is only one vertex, and the whole strategy is determined by identifying who is winning (whether this is the player on move or the other one) Directed edges are used for the description of possible moves and there is then always an acyclic graph.



The example of the game Nim is displayed on the diagram. On the left, there is a complete game tree corresponding to three tokens. The directed graph on the right represents the game with seven tokens. A complete tree for this game would already have 21 leaves, and the number of leaves grows exponentially with the number of tokens.

We thus get the following recurrent formula for the expected number of comparisons:

$$C_n = n - 1 + \sum_{k=1}^n \frac{1}{n} (C_{k-1} + C_{n-k}).$$

It is possible to solve this recurrence (using certain tricks which can be learned to some extent) even without using generating functions.

$$C_n = n - 1 + \frac{2}{n} \sum_{k=1}^n C_{k-1}$$

symmetry of both

$$nC_n = n(n - 1) + 2 \sum_{k=1}^n C_{k-1}$$

multiply by  $n$

$$(n - 1)C_{n-1} = (n - 1)(n - 2) + 2 \sum_{k=1}^{n-1} C_{k-1}$$

the same expression

$$nC_n = (n + 1)C_{n-1} + 2(n - 1)$$

subtracted and rearranged

We have thus obtained a much simpler recurrence:

$$nC_n = (n + 1)C_{n-1} + 2(n - 1).$$

On the other hand, this equation contains non-constant coefficients as well.

We can also note that the recurrence has been simplified to the extent that the values  $C_n$  can be computed iteratively. Nevertheless, it is advantageous to express these values explicitly as a function of  $n$  (or at least to approximate them).

First, we use a slight trick: dividing both sides by  $n(n + 1)$ :

$$\frac{C_n}{n + 1} = \frac{C_{n-1}}{n} + \frac{2(n - 1)}{n(n + 1)}$$

Now, we “expand” this expression (*telescope*, we can also use the substitution  $B_n = C_n/n + 1$ ):

$$\frac{C_n}{n + 1} = \frac{2(n - 1)}{n(n + 1)} + \frac{2(n - 2)}{(n - 1)n} + \dots + \frac{2 \cdot 1}{2 \cdot 3} + \frac{C_1}{2}.$$

Hence

$$\frac{C_n}{n + 1} = 2 \sum_{k=1}^{n-1} \frac{k}{(k + 1)(k + 2)}.$$

This can be summed up using partial fraction decomposition, for instance:  $\frac{k}{(k+1)(k+2)} = \frac{2}{k+2} - \frac{1}{k+1}$ . This leads to

$$\frac{C_n}{n + 1} = 2 \left( H_{n+1} - 2 + \frac{1}{n + 1} \right),$$

whence

$$C_n = 2(n + 1)H_{n+1} - 4(n + 1) + 2$$

( $H_n = \sum_{k=1}^n \frac{1}{k}$  is the sum of the first  $n$  terms of the harmonic progression). At the same time, we can give the bound  $H_n \sim \int_1^n \frac{dx}{x} + \gamma$ , whence

$$C_n \sim 2(n + 1)(\ln(n + 1) + \gamma - 2) + 2.$$

**13.1.11.** Using the generating function  $F(x) = x/(1 - x - x^2)$  for the Fibonacci sequence, find the generating function for the “semi-Fibonacci” sequence  $(F_0, F_2, F_4, \dots)$ .  $\circ$

The individual vertices in the directed acyclic graph on the right-hand side of the diagram indicate the number of tokens left and the information whether the game at that state is won by the player who is to move (letter  $N$  as “next”) or the other one (letter  $P$  as “previous”). Altogether, considering a game with  $k$  tokens, this graph always has only  $k + 1$  vertices. At the same time, there is the complete strategy encoded in the graph: The players always try to move from the current state into a vertex labeled by  $P$  if such one exists.

In fact, every directed acyclic graph can be seen as a description of a game. The initial situations are represented by those vertices which have no incoming edges (there can be none or more of them), and the game ends in leaves, i.e. vertices with no outgoing edges (again, there can be one or more of them).

The strategy for each player can be obtained by a simple recursive procedure as above (for the sake of simplicity, it is assumed that there is no tie):

- The leaves are labeled by  $P$  ( the player who is to move from a leaf loses).
- A non-leaf vertex of the graph is labeled by  $N$  if there is an edge leading to a  $P$ -vertex. Otherwise, it is labeled  $P$ .

In the case of our variation of Nim, the situation is very simple. It follows from the strategy described that the player who is to move loses if and only if the number of tokens is divisible by three.

The games that can be represented by a directed acyclic graph are called *impartial*. These are exactly those games which satisfy:

- in every state, both players choose from the same set of moves;
- the number of possible states is finite;
- the game has “zero sum”, i.e. the better the outcome for one of the players, the worse for the other one.

An example of an impartial game is tic-tac-toe. Although the players use different symbols in this game, they can place them in any of the unoccupied squares. On the other hand, chess is not an impartial game in this sense, since the set of possible moves in every situation depends on the number of pieces the players have at their disposal.

**13.2.16. Sum of combinatorial games.**



The rules of the real classical game Nim are somewhat more complicated: There are three piles of tokens. The move consists of selecting one of the piles and removing an arbitrary (positive) number of tokens from that pile. The player who manages to take the last token wins. There is a variation of the game in which the player who is forced to take the last token loses. If this game is considered with one pile, the situation is easy: The first player takes all the tokens and wins immediately. However, it is not that easy with three piles. Whether the analysis of the one-pile game is of any use for this more complicated game is a good question.

**13.I.12.** The fan of order  $n$  is a graph on  $n + 1$  vertices, which are labeled  $0, 1, \dots, n$ , with the following edges: vertex  $0$  is connected to all other vertices, and for each  $k$  satisfying  $1 \leq k < n$ , vertex  $k$  is connected to vertex  $k + 1$ . How many spanning trees does this graph have?

**Solution.** Denoting by  $v_n$  the wanted number of spanning trees, we clearly have  $v_1 = 1$ , and since the fan of order 2 is the triangle graph  $K_3$ , we have  $v_2 = 3$ . Further, we are going to show that for  $n > 1$ , the following recurrence<sup>3</sup> holds:

$$v_n = v_{n-1} + \sum_{k=0}^{n-1} v_k + 1, \quad v_0 = 0.$$

For a fixed spanning tree of the fan of order  $n$ , let  $k$  be the greatest integer of the set  $\{1, \dots, n - 1\}$  such that the spanning tree contains all edges of the path  $(0, 1, 2, 3, \dots, k)$ . This spanning tree cannot contain the edges  $\{0, 2\}, \dots, \{0, k\}, \{k, k + 1\}$ ; therefore, there are the same number of spanning trees for a fixed  $k$  as in the fan of order  $n - k$  with vertices  $0, k + 1, k + 2, \dots, n$ , i. e.  $v_{n-k}$ . Further, we must count one spanning tree for  $k = n$  and those spanning trees which do not contain the edge  $\{0, 1\}$  (thus they must contain the edge  $\{1, 2\}$ ) – they are obtained from fans of order  $n - 1$  on vertices  $0, 2, \dots, n$ . We have thus obtained the wanted recurrence  $v_n = v_{n-1} + v_{n-1} + v_{n-2} + \dots + v_0 + 1$ .

Now, we have the general formula

$$v_n = v_{n-1} + \sum_{k=0}^{n-1} v_k + 1 - [n = 0],$$

whence the usual procedure for finding the generating function  $V(x)$  of this sequence yields

$$\begin{aligned} V(x) &= x \cdot V(x) + \sum_{n=0}^{\infty} \sum_{k < n} v_k x^n + \frac{1}{1-x} - 1 = \\ &= x \cdot V(x) + \sum_{k=0}^{\infty} \sum_{n > k} v_k x^n + \frac{x}{1-x} = \\ &= \left( \sum_{k=0}^{\infty} v_k x^k \right) \cdot \sum_{n > k} x^{n-k} + \frac{x}{1-x} = \\ &= \left( \sum_{k=0}^{\infty} v_k x^k \right) \cdot \frac{x}{1-x} + \frac{x}{1-x} = V(x) \cdot \frac{x}{1-x} + \frac{x}{1-x} \end{aligned}$$

The solution of the equation  $V(x) = xV(x) + \frac{x}{1-x}V(x) + \frac{x}{1-x}$  is

$$V(x) = \frac{x}{1 - 3x + x^2},$$

whence using the standard method (partial fraction decomposition) or the previous problem leads to the result  $v_n = F_{2n}$ .  $\square$

<sup>3</sup> Using this recurrent formula to calculate more values  $v_n$ , we find out that  $v_3 = 8, v_4 = 21$ , which suggests a hypothesis about connection with the Fibonacci sequence in the form  $v_n = F_{2n}$ . This can be proved easily by induction.

For this purpose, introduce a new concept, the *sum of impartial games*: A situation in the game composed of two simpler games is a pair of possible situations in the particular games. Then, a move consists of selecting one of the two games and performing a move in that game (the other game is left unchanged). Therefore, the sum of impartial games is an operation which assigns to a pair of directed acyclic graphs a new one.

Considering graphs  $G_1 = (V_1, E_1)$  and  $G_2 = (V_2, E_2)$ , its sum  $G_1 + G_2$  is the graph  $G = (V, E)$ , where  $V = V_1 \times V_2$  and

$$E = \{(v_1 v_2, w_1 v_2); (v_1, w_1) \in E_1\} \cup \{(v_1 v_2, v_1 w_2); (v_2, w_2) \in E_2\}.$$

In the case of one game, the vertices can be labeled step-by-step by the letters  $N$  and  $P$  in an upwards manner, according to whether one can get to a  $P$ -vertex along some of the edges. However, in the sum of games, movement along the edges is needed in a much more complicated way. Therefore, finer tools are needed for expressing the reachability of vertices labeled by  $P$  from other vertices.

This needs some preparation which might seem like a strange magic (but the proof of the theorem below shows that all this is quite natural). Define the *Sprague–Grundy function* recursively.  $g : V \rightarrow \mathbb{N}$  on a directed acyclic graph  $G = (V, E)$  as follows:<sup>11</sup>

- (1) for a leaf  $v$ , set  $g(v) = 0$ ;
- (2) for a non-leaf vertex  $v \in V$ , define

$$g(v) = \text{mex}\{g(w); (v, w) \text{ is an edge}\},$$

where the *minimum excluded value* function  $\text{mex}$  is defined on subsets  $S$  of the natural numbers  $\mathbb{N} = \{0, 1, \dots\}$  by

$$\text{mex } S = \min \mathbb{N} \setminus S.$$

The function  $g(v)$  is just the  $\text{mex } S$  operation for the set  $S$  of the values  $g(w)$  over those vertices  $w$  where one can get along an edge from  $v$ .

Note that this definition is correct since, clearly, the formula uniquely defines a function that assigns a natural number to any vertex in the acyclic graph in question.

Yet another operation on the natural numbers is needed. It is the binary XOR operation

$$(a, b) \mapsto a \oplus b,$$

performing the exclusive-or operation bit-wise on the binary expansions of  $a$  and  $b$ . This operation can be considered from the following point of view: Consider the binary expansions of  $a$  and  $b$  to be vectors in the vector space  $(\mathbb{Z}_2)^k$  over  $\mathbb{Z}_2$  (for a sufficiently large  $k$ ), and add them there. The resulting vector is the binary expansion of  $a \oplus b$ .

Now the main result can be formulated:

<sup>11</sup>We are presenting the theory which was developed in combinatorial game theory independently by R. P. Sprague in 1935 and P. M. Grundy in 1939.

**Recursively connected sequences.** Sometimes, we are able to express the wanted number of ways or events only in terms of more mutually connected sequences.

**13.I.13.** In how many ways can we cover a  $3 \times n$  rectangle with  $1 \times 2$  domino pieces? Evaluate this value for  $n = 20$ .

**Solution.** We can easily find out that  $c_1 = 0, c_2 = 3, c_3 = 0$ , and it is reasonable to set  $c_0 = 1$  (this is not merely convention; there is indeed a unique empty covering).

We are looking for a recursive formula—discussing the behavior “on the edge”, we find out that  $c_n = 2r_{n-1} + c_{n-2}$ ,  $r_n = c_{n-1} + r_{n-2}$ ,  $r_0 = 0, r_1 = 1$ , where  $r_n$  is the number of coverings of the rectangle  $3 \times n$  without one of the corner tiles.

The values of  $c_n$  and  $r_n$  for the first few non-negative integers  $n$  are:

$n$	0	1	2	3	4	5	6	7
$c_n$	1	0	3	0	11	0	41	0
$r_n$	0	1	0	4	0	15	0	56

- Step 1:  $c_n = 2r_{n-1} + c_{n-2} + [n = 0]$ ,  $r_n = c_{n-1} + r_{n-2}$ .
- Step 2:

$$C(x) = 2xR(x) + x^2C(x) + 1, \quad R(x) = xC(x) + x^2R(x).$$

- Step 3:

$$C(x) = \frac{1 - x^2}{1 - 4x^2 + x^4}, \quad R(x) = \frac{x}{1 - 4x^2 + x^4}.$$

- Step 4: We can see that both are functions of  $x^2$ . We can thus save much work if we consider the function  $D(z) = 1/(1 - 4z + z^2)$ . Then, we have  $C(x) = (1 - x^2)D(x^2)$ , i. e.,  $[x^{2n}]C(x) = [x^{2n}](1 - x^2)D(x^2) = [x^n](1 - x)D(x)$ , so  $c_{2n} = d_n - d_{n-1}$ .

The roots of  $1 - 4x + x^2$  are  $2 + \sqrt{3}$  and  $2 - \sqrt{3}$ , whence the standard procedure yields

$$c_{2n} = \frac{(2 + \sqrt{3})^n}{3 - \sqrt{3}} + \frac{(2 - \sqrt{3})^n}{3 + \sqrt{3}}.$$

Just like with the Fibonacci sequence, the second term is negligible for large values of  $n$  and is always between 0 and 1. Therefore,

$$c_{2n} = \left\lceil \frac{(2 + \sqrt{3})^n}{3 - \sqrt{3}} \right\rceil.$$

For instance,  $c_{20} = 413403$ . □

**13.I.14.** Using generating functions, find the number of ones in a random bit string.

**Solution.** Let  $B$  the set of bit strings, and for  $b \in B$ , let  $|b|$  denote the length of  $b$  and  $j(b)$  the number of ones in it. The generating function is of the form

$$B(x) = \sum_{b \in B} x^{|b|} = \sum_{n \geq 0} 2^n x^n = \frac{1}{1 - 2x}.$$

SPRAGUE–GRUNDY THEOREM

**13.2.17. Theorem.** Consider a directed acyclic graph  $G = (V, E)$ . Its vertices  $v$  are labeled by  $P$  if and only if  $g(v) = 0$ , where  $g$  is the Sprague–Grundy function.

For any two directed acyclic graphs  $G_1 = (V_1, E_1)$ ,  $G_2 = (V_2, E_2)$  and their Sprague–Grundy functions  $g_1, g_2$ , the Sprague–Grundy function  $g$  of their sum is given by

$$g(v_1 v_2) = g_1(v_1) \oplus g_2(v_2).$$



**PROOF.** The first proposition of the theorem follows directly by induction from the definition of the Sprague–Grundy function  $g$ .

The proof of the other part is more complicated. Let  $(v_1 v_2)$  be a position of the game  $G_1 + G_2 = (V, E)$ , and consider any  $a \in \mathbb{N}_0$  such that  $a < g_1(v_1) \oplus g_2(v_2)$ . There exists a state  $x_1 x_2$  of the game  $G_1 + G_2$  such that  $g(x_1) \oplus g(x_2) = a$  and  $(v_1 v_2, x_1 x_2) \in E$ , and, at the same time, there is no edge  $(v_1 v_2, y_1 y_2) \in E$  such that

$$g_1(y_1) \oplus g_2(y_2) = g_1(v_1) \oplus g_2(v_2).$$

This justifies the recursive definition of the Sprague–Grundy function and proves the rest of the theorem.

To show why, find a vertex  $x_1 x_2$  with a given value  $a < g_1(v_1) \oplus g_2(v_2)$  of the Sprague–Grundy function.

Consider the integer  $b := a \oplus g_1(v_1) \oplus g_2(v_2)$ . Refer to the bit of value  $2^i$  as the  $i$ -th bit of an integer. Clearly,  $b \neq 0$ . Let  $k$  be the position of the highest one in the binary expansion of  $b$ , i.e.  $2^k \leq b < 2^{k+1}$ . This means that the  $k$ -th bit of exactly one of the integers  $a, g_1(v_1) \oplus g_2(v_2)$  is one and that these integers do not differ in higher bits. It follows from the assumption  $a < g_1(v_1) \oplus g_2(v_2)$  that it is the integer  $g_1(v_1) \oplus g_2(v_2)$  whose  $k$ -th bit is one.

Therefore, the  $k$ -th bit of exactly one of the integers  $g_1(v_1), g_2(v_2)$  is one. Assume without loss of generality that it is the integer  $g_1(v_1)$ . Further, consider the integer  $c := g_1(v_1) \oplus b$ . Recall that the highest one of  $b$  is at position  $k$ , so the integers  $c, g_1(v_1)$  do not differ in higher bits and the  $k$ -th bit of  $c$  is zero. Therefore,  $c < g_1(v_1)$ . Then, by the definition of the function value  $g_1(v_1)$ , there must exist a state  $w_1$  of the game  $G_1$  such that  $(v_1, w_1) \in E_1$  and  $g_1(w_1) = c$ . Now,  $(v_1 v_2, w_1 v_2) \in E$  and

$$\begin{aligned} g_1(w_1) \oplus g_2(v_2) &= c \oplus g_2(v_2) = g_1(v_1) \oplus b \oplus g_2(v_2) \\ &= g_1(v_1) \oplus a \oplus g_1(v_1) \oplus g_2(v_2) \oplus g_2(v_2) = a. \end{aligned}$$

This fulfills the first part of our plan.

Further, consider any edge  $(v_1 v_2, y_1 y_2) \in E$  in  $G$ , where  $(v_1, y_1) \in E_1$ , and hence  $v_2 = y_2$ . Suppose that  $g_1(y_1) \oplus g_2(y_2) = g_1(v_1) \oplus g_2(v_2)$ . Then,  $g_1(y_1) \oplus g_2(y_2) = g_1(v_1) \oplus g_2(v_2)$ . Clearly, the terms  $g_2(v_2)$  can be canceled (it is an operation in a vector space), leading to  $g_1(y_1) = g_1(v_1)$ . This contradicts the properties of the Sprague–Grundy function  $g_1$  of the game  $G_1$ . This proves the second part of the theorem. □



The generating function for the number of ones is

$$C(x) = \sum_{b \in B} j(b)x^{|b|}.$$

A string  $b$  can be obtained from the one bit shorter string  $b'$  by adding either a zero or a one, i. e.,  $j(b)$  is the sum of  $j(b')$  ones and  $j(b') + 1$  ones. Therefore,

$$\begin{aligned} C(x) &= \sum_{b' \in B} (1 + 2j(b'))x^{|b'|+1} = \sum_{b' \in B} x^{|b'|+1} + 2 \sum_{b' \in B} j(b')x^{|b'|+1} \\ &= xB(x) + 2xC(x). \end{aligned}$$

Hence

$$C(x) = \frac{x}{(1-2x)^2} = x(1-2x)^{-2}$$

and the  $n$ -th coefficient is  $c_n = 2^{n-1} \binom{-2}{n-1} = n2^{n-1}$ . This number gives the number of ones in strings of length  $n$ , and there are  $b_n = 2^n$  such strings. Therefore, the expected number of ones in such string is  $\frac{c_n}{b_n} = \frac{n}{2}$ , which is, of course, what we have anticipated.  $\square$

**13.I.15.** Find the generating function and an explicit formula for the  $n$ -th term of the sequence  $\{a_n\}$  defined by the recurrent formula

$$\begin{aligned} a_0 &= 1, a_1 = 2 \\ a_n &= 4a_{n-1} - 3a_{n-2} + 1 \text{ for } n \geq 2. \end{aligned}$$

**Solution.** The universal formula which holds for all  $n \in \mathbb{Z}$  is

$$a_n = 4a_{n-1} - 3a_{n-2} + 1 - 3[n = 1].$$

Multiplying by  $x^n$  and summing over all  $n$ , we get the following equation for the generating function:  $A(x) = \sum_{n=0}^{\infty} a_n x^n$ .

Hence, we can express

$$A(x) = \frac{3x^2 - 3x + 1}{(1-x)^2(1-3x)} = \frac{3}{4} \cdot \frac{1}{1-x} - \frac{1}{2} \cdot \frac{1}{(1-x)^2} + \frac{3}{4} \cdot \frac{1}{1-3x}$$

Therefore, the coefficient at  $x^n$  is

$$\begin{aligned} a_n &= \frac{3}{4}(-1)^k \binom{-1}{n} - \frac{1}{2}(-1)^n \binom{-2}{n} + \frac{3}{4}(-3)^n \binom{-1}{n} \\ &= \frac{3}{4} - \frac{1}{2}(n+1) + \frac{3}{4}3^n = d \frac{1-2n+3^{n+1}}{4}. \quad \square \end{aligned}$$

**13.I.16.** Solve the following recurrence using generating functions:

$$\begin{aligned} a_0 &= 1, a_1 = 2 \\ a_n &= 5a_{n-1} - 4a_{n-2} \quad n \geq 2 \end{aligned}$$

**Solution.** The universal formula is of the form

$$a_n = 5a_{n-1} - 4a_{n-2} - 3[n = 1] + [n = 0]$$

Multiplying both sides by  $x^n$  and summing over all  $n$ , we obtain

$$A(x) = 5xA(x) - 4x^2A(x) - 3x + 1.$$

Hence

$$A(x) = \frac{1-3x}{(1-4x)(1-x)} = \frac{2}{3} \cdot \frac{1}{1-x} + \frac{1}{3} \cdot \frac{1}{1-4x}$$

The following useful result is a direct corollary of this theorem:

**Corollary.** A vertex  $v_1v_2$  in the sum of games is labeled by  $P$  if and only if  $g_1(v_1) = g_2(v_2)$ .

For example, if three piles of tokens are combined in the simplified Nim game (it is always allowed to take only one or two tokens), the first player always wins, if all three piles have the same number of tokens, not divisible by three. The individual functions  $g_i(k)$  for the individual piles equal the remainder after dividing  $k$  by 3. It follows that, when summing the first two Nim pile games, the value  $g(v) = 0$  is obtained for the initial position. Summing again with another pile game gives  $g(v) \neq 0$ .

In the original game, the individual piles are described by  $g(k) = k$  (any number of tokens can be chosen, hence the function  $g$  grows in this way). The losing positions are those, where the binary sum of the numbers of tokens is zero. For example, if the two of the initial piles are of equal size, then a simple winning strategy is to remove the third one completely and always make the remaining two equal after the opponent's move.

**Remark.** Further details are omitted in this text. It can be proved that every finite directed acyclic graph is isomorphic to a finite sum of suitably generalized games of Nim.

In particular, the analysis of the simple game and construction of the function  $g$  basically (at least implicitly) gives a complete analysis of all impartial games.

### 3. Remarks on Computational Geometry

A large amount of practical problems consist in constructing or analyzing some finite geometrical objects in Euclidean spaces, mainly in 2D or 3D. This is a very busy area of both applications and research. At the same time, most of the algorithms and their complexity analysis are based on graph theoretical and further combinatorial concepts. We provide several glimpses into this beautiful topic.<sup>12</sup> We discuss convex hulls, triangulations, and Voronoi diagrams and focus on a few basic approaches only.

**13.3.1. Convex hulls.** We start with a simple and practical problem. In the plane  $\mathbb{R}^2$ , suppose  $n$  points  $X = \{v_1, \dots, v_n\}$  are given and the task is to find their convex hull  $CH(X)$ . As learned in the Chapter 4,  $CH(X)$  is given by a convex polygon and it is desired to find it effectively.

First we have to decide how  $CH(X)$  should be encoded as a data structure. Choose the connected list of edges. There is the cyclic list of the vertices  $v_i$  in the polygon, sorted in the counter clock-wise order, together with pointers towards the oriented segments between the consecutive vertices (the

<sup>12</sup>The beautiful book *Computational Geometry, Algorithms and Applications* by de Berg, M., Cheong, O., van Kreveld, M., Overmars, M., published by Springer (1997) can be warmly recommended, <http://www.springer.com/us/book/9783540779735>

and

$$a_n = \frac{2}{3} \binom{-1}{n} + \frac{2}{3} \binom{-1}{n} (-4)^n = \frac{4^n + 2}{3}. \quad \square$$

**13.I.17.** A cash dispenser can provide us with banknotes of values 200, 500, and 1,000 crowns. In how many ways can we pick 7,000 crowns? Use generating functions to find the solution.

**Solution.** The problem can be reformulated as looking for the number of integer solutions of the equation

$$2a + 5b + 10c = 70; \quad a, b, c \geq 0.$$

This number is equal to the coefficient at  $x^{70}$  in the function

$$G(x) = (1+x^2+x^4+\dots)(1+x^5+x^{10}+\dots)(1+x^{10}+x^{20}+\dots)$$

This function is equal to

$$G(x) = \frac{1}{1-x^2} \frac{1}{1-x^5} \frac{1}{1-x^{10}}$$

and since

$$\frac{1-x^{10}}{1-x^5} = 1+x^5 \text{ and } \frac{1-x^{10}}{1-x^2} = 1+x^2+x^4+x^6+x^8,$$

we can transform it into the form

$$G(x) = \frac{(1+x^2+x^4+x^6+x^8)(1+x^5)}{(1-x^{10})^3}.$$

By the binomial theorem, we have

$$(1-x^{10})^3 = \sum_{k=0}^{\infty} (-1)^k \binom{-3}{k} x^{10k}.$$

Therefore,  $G(x)$  equals

$$(1+x^2+x^4+x^5+x^6+x^7+x^8+x^9+x^{11}+x^{13}) \sum_{k=0}^{\infty} (-1)^k \binom{-3}{k} x^{10k}$$

The term  $x^{70}$  can be obtained only as  $7 \cdot 10 + 0$ , i. e., the coefficient at  $x^{70}$  is equal to

$$[x^{70}]G(x) = -\binom{-3}{7} = \binom{3+7-1}{7} = \frac{9 \cdot 8}{2} = 36. \quad \square$$

**13.I.18.** Find the probability of getting exactly  $k$  heads after  $n$  times tossing the coin.

**Solution.** Represent  $H$  outcome as  $x$  and  $T$  as  $y$ . We see that all possible outcomes of  $n$  tosses are represented by expansion of  $f(x) = (x+y)^n$ . The coefficient at  $x^k y^{n-k}$  is the number of outcomes with exactly  $k$  heads. The required probability is, therefore,  $\frac{\binom{n}{k}}{2^n}$ .  $\square$

**13.I.19.** Find the probability of getting exactly  $k$  heads after  $n$  times tossing the coin if the probabilities of a head and a tail equal  $p$  and  $q$ , respectively,  $P + q = 1$ .

**Solution.** Represent  $H$  outcome as  $px$  and  $T$  as  $q$ . We see that all possible outcomes of  $n$  tosses are represented by expansion of  $f(x) = (px + q)^n$ . The coefficient at  $x^k$  is the number of outcomes with exactly  $k$  heads. The required probability is, therefore,  $\frac{p^k q^{n-k} \binom{n}{k}}{2^n}$ .  $\square$

edges). Moreover, there is the list of edges pointing to their tail and head vertices.

There is a simple way, to get  $CH(X)$ . Namely, create the oriented edges  $e_{ij}$  for all pairs  $(v_i, v_j)$  of the points in  $X$ , and decide whether  $e_{ij}$  belongs to  $CH(X)$  by testing whether all the other points of  $X$  are on the left of  $e_{ij}$  (in the obvious sense). It is known already from chapter one, that this is tested in constant time by means of the determinant. Clearly  $e_{ij}$  belongs to  $CH(X)$  if and only if all the latter tests are positive. In the end, the order in which to sort the edges and vertices in the output is found.

This does not look like a good algorithm, since  $O(n^2)$  edges need to be tested against  $O(n)$  points. Hence, cubic time complexity is expected. But there is a strong simple improvement available. Consider the lexicographic order on the points  $v_i$  with respect to their coordinates. Then build the convex hull consecutively and run through the tests only for the edges having the last added vertex as their tail.

GIFT WRAPPING CONVEX HULL ALGORITHM

*Input:* A set of points  $X = \{v_1, \dots, v_n\}$  in the plane.

*Output:* The requested edge list for  $CH(X)$ .

- (1) *Initialization.* Take the smallest vertex  $v_0$  in the lexicographic order with respect to the coordinates, and set  $v_{\text{active}} = v_0$ .
- (2) *Main cycle.*
  - Test edges with tail  $v_{\text{active}}$ , until  $e$  belonging to  $CH(X)$  is found.
  - add  $e$  to  $CH(X)$  and set its head to be the  $v_{\text{active}}$
  - if  $v_{\text{active}} \neq v_0$ , then repeat the cycle.

Obviously, the most left and lowest vertex  $v_0$  in  $X$  is in  $CH(X)$ . Since  $CH(X)$  is a cycle (as a directed graph), the algorithm works correctly. It is necessary to be careful about possible collinear edges in  $CH(X)$  and the lack of robustness of the test for those nearly collinear.



This simple improvement reduces the worst running time of the algorithm to  $O(n^2)$ . The worst case can be obtained if all the points  $v_i$  appear on one circle, and unluckily the right next point is always found as the very last one in partial tests. But the actual running time is much better, at most  $O(ns)$ , where  $s$  is the size of the  $CH(X)$ .

For example, in situations where the distribution of the points in the plane is random with normal distribution (see

**13.I.20.** There are coins  $C_1, \dots, C_n$ . For each  $k$  coin  $C_k$  is biased so that when tossed the probability that it falls heads is  $\frac{1}{2k+1}$ . If  $n$  coins are tossed, what is the probability of an odd number of heads?

**Solution.** Let  $p_k = \frac{1}{2k+1}$  and  $q_k = 1 - p_k$ . The generating function can be written as

$$f(x) = \prod_{i=1}^n (q_i + p_i x) = \sum_{m=0}^n a_m x^m,$$

with  $a_m$  the probability of getting exactly  $m$  heads. Observe that

$$f(1) = \sum a_m = \sum a_{2k} + \sum a_{2k+1}.$$

and

$$f(-1) = \sum a_m (-1)^m = \sum a_{2k} - \sum a_{2k+1}$$

. Hence,

$$\begin{aligned} \sum a_{2k+1} &= \frac{1}{2}(f(1) - f(-1)) = \frac{1}{2}\left(1 - \frac{1}{2n+1}\right) = \frac{n}{n+1}, \\ \text{as } f(1) &= 1 \text{ and } f(-1) = \frac{1}{3} \frac{3}{5} \frac{5}{7} \dots \frac{2n-1}{2n+1} = \frac{1}{2n+1}. \quad \square \end{aligned}$$

**13.I.21.** Let  $n$  be a positive integer. Find the number of polynomials  $P(x)$  with coefficients from the set  $\{0, 1, 2, 3\}$  such that  $P(2) = n$ .

**Solution.** Let  $P(x) = a_0 + a_1x + a_2x^2 + \dots + a_kx^k + \dots$  is such polynomial. Then its coefficients should satisfy the equation

$$a_0 + 2a_1 + 4a_2 + \dots + 2^k a_k + \dots = n$$

with integer coefficients  $0 \leq a_k \leq 3$ . The number of such solutions is given by the coefficient at  $x^n$  in

$$\begin{aligned} f(x) &= (1+x+x^2+x^3)(1+x^2+x^4+x^6)(1+x^4+x^8+x^{12}) \dots \\ &= \prod_{k=0}^{\infty} (1+x^{2^k} + x^{2(2^k)} + x^{3(2^k)}) \\ &= \prod_{k=0}^{\infty} \frac{1-x^{2^{k+2}}}{1-x^{2^k}} = \frac{1}{(1-x)(1-x^2)} \\ &= \frac{1}{4(1-x)} + \frac{1}{2(1-x)^2} + \frac{1}{4(1+x)} = \sum_{n=0}^{\infty} \left(\frac{1}{4} + (-1)^n \frac{1}{4} + \frac{1}{2(n+1)}\right) x^n \end{aligned}$$

Hence, the required coefficient is  $\lfloor \frac{n}{2} \rfloor + 1$ .  $\square$

**13.I.22.** Find the number of subsets of  $\{1, \dots, 2017\}$ , such that the sum of its elements is divisible by 5.

**Solution.** For subset  $S = \{s_1, s_2, \dots, s_k\}$  define  $\sigma(S) = x^{s_1} x^{s_2} \dots x^{s_k}$ . Then

$$\sum_{S \subset \{1, \dots, 2017\}} \sigma(S) = (1+x)(1+x^2) \dots (1+x^{2017}).$$

The coefficient at  $x^n$  in  $f(x)$  is the number of subsets in  $\{1, \dots, 2017\}$  with the sum of its elements being exactly  $n$ . Let  $f(x) = \sum a_n x^n$ . We are looking for the sum  $A_5 = a_0 + a_5 + a_{10} + \dots$ . Let  $\omega_j = \exp(\frac{2j\pi}{5}), j = 0, \dots, 4$  be the 5-th roots of unity with  $\omega_0 = 1$ . Then

$$A_5 = \frac{1}{5}(f(\omega_0) + f(\omega_1) + f(\omega_2) + f(\omega_3) + f(\omega_4)).$$

the chapter 10 for what this means), then it is known that the expected size would be logarithmic.

At the same time, finding  $CH(X)$  for  $X$  distributed on a circle is equivalent to sorting the points along the circle. So the worst time run cannot be better than  $O(n \log n)$  for all algorithms (cf. 13.1.17).

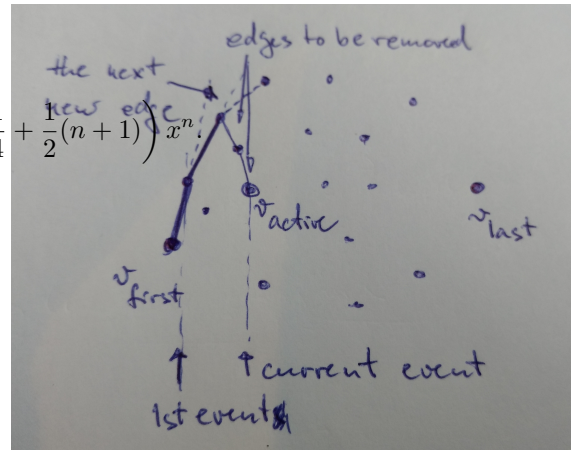
**13.3.2. The sweep line paradigm.** We illustrate several main approaches to computational geometry algorithms for the same convex hull problem.



The latter algorithm is close to the idea of having a special object  $L$  running through all the objects in the input  $X$  consecutively, taking care of constructing the relevant partial output of the algorithm on the run. This is the *event structure* describing all the events needing consideration, and the *sweep line structure* carrying all the information to deal with the individual events. The procedure is similar to the search over a graph discussed earlier. For a shortcut, this may reduce the dimension of the problem (e.g. from 2D to 1D) on the cost of implementing dynamical structures.

To start, initialize the queue of the events and begin dealing with them. At each step, there are the still sleeping events (those further in the queue not yet treated), the current active events (those under consideration) and the already processed ones.

With the  $CH(X)$ , this idea can be implemented as follows. Initialize the lexicographically ordered points  $v_i$  in  $X$ . Notice that the first and last ones necessarily appear in the  $CH(X)$ . This way,  $CH(X)$  splits into the two disjoint chains of edges between them. We call them the upper and the lower convex hulls. Hence the entire construction can be split into the upper convex hull and lower convex hull.



As in the diagram, moving through the events one by one, it is only needed to check, whether the edge joining the last but one active vertex with the current one makes a left or right turn (as usual, right means the clockwise orientation). If right, then add the edge to list, if left, omit the recent edges one by one, until the right turn is obtained.

Obviously  $f(\omega_0) = f(1) = 2^{2017}$ . For  $1 \leq j \leq 4$

$$f(\omega_j) = ((1+\omega_j^0)(1+\omega_j)(1+\omega_j^2)(1+\omega_j^3)(1+\omega_j^4))^{403}(1+\omega_j)(1+\omega_j^2)(1+\omega_j^3)(1+\omega_j^4)$$

Therefore,

$$\sum_{j=1}^4 F(\omega_j) = 2^{403}(4 - 1 - 1 - 1) = 2^{403}.$$

Therefore, the answer is  $\frac{1}{5}(2^{2017} + 2^{403})$ , which is an integer as 2 is the last digit in  $2^{2017}$  and 8 is the last digit in  $2^{403}$ .  $\square$

**13.I.23.** Using generating functions, solve recursive equation

$$a_{k+1} = 2a_k + 4^k, \quad a_1 = 3.$$

**Solution.** From  $a_1 = 2a_0 + 4^0$  follows  $a_0 = 1$ . Multiplying both sides of the recursion by  $x^k$  by summation on  $k$  we obtain

$$\sum_{k=0}^{\infty} a_{k+1}x^k = 2 \sum_{k=0}^{\infty} a_kx^k + \sum_{k=0}^{\infty} 4^kx^k.$$

Therefore, for generating function  $f(x) = \sum_{k=0}^{\infty} a_kx^k$  we have

$$\frac{1}{x} \sum_{k=0}^{\infty} a_{k+1}x^{k+1} = 2f(x) + \sum_{k=0}^{\infty} (4x)^k,$$

that is

$$\frac{1}{x}(f(x) - 1) = 2f(x) + \frac{1}{1 - 4x}.$$

Therefore,

$$f(x) = \frac{1}{1 - 2x} + \frac{x}{(1 - 2x)(1 - 4x)} = \frac{\frac{1}{2}}{1 - 2x} + \frac{\frac{1}{2}}{1 - 4x}.$$

Expanding  $f(x)$ , we obtain:

$$f(x) = \frac{1}{2} \sum_{k=0}^{\infty} (2x)^k + \frac{1}{2} \sum_{k=0}^{\infty} (4x)^k,$$

and so,  $a_k = \frac{1}{2}2^k + \frac{1}{2}4^k$ .  $\square$

**13.I.24.** In how many ways can  $n$  balls be distributed in 4 boxes if the first box has at least two balls.

**Solution.** The generating function for the first box is  $b_1(x) = x^2 + x^3 + x^4 + \dots = \frac{x^2}{1-x}$ . The generating function for each other box is  $b(x) = 1 + x + x^2 + x^3 + \dots = \frac{1}{1-x}$ . For all 4 boxes it has to be  $f(x) = b_1(x)(b(x))^3 = \frac{x^2}{(1-x)^4}$ . The coefficient at  $x^n$  in  $f(x)$  that indicates the number of options to distribute  $n$  balls, equals

$$\binom{n}{n} - 2 + 4 - 1, 4 - 1 = \binom{n}{n} + 1, 3).$$

SWEEP LINE UPPER CONVEX HULL

**Input:** A set of points  $X = \{p_1, \dots, p_n\}$  in the plane.  
**Output:** The directed path  $UCH(X)$ .

(1) *Initialization.*

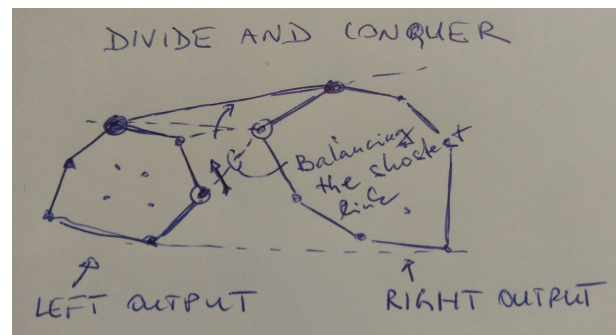
- Set the event structure to be the lexicographically ordered list of points  $v_{\text{first}}, \dots, v_{\text{last}}$ . There is no special sweep line structure but the indicator distinguishing the stage of the event.
- Set the active event to be  $v_{\text{active}} = v_{\text{first}}$  and initiate the  $UCH(X)$  as the trivial path with one vertex  $v_{\text{first}}$  (this is the current last vertex of the path in construction).

(2) *Main cycle.*

- Set the active event to the next point  $v$  in the queue, and consider the potential edge  $e$  having  $v_{\text{active}}$  as the tail and the last vertex in  $UCH(X)$  as head.
- Check whether the  $UCH(X)$  is to the left of  $e$  (check it only against the last edge in the current  $UCH(X)$ ). If so, add  $e$  and the  $v_{\text{active}}$  to the  $UCH(X)$ . If not, remove edges in  $UCH(X)$  one by one, until the test turns positive.
- Repeat the cycle until the next event is the  $v_{\text{last}}$ .

It is easy to check that the algorithm is correct. Exactly  $n$  events need to be considered, and at each of it up to  $O(n)$  vertices can be removed in the current  $UCH(X)$ . This occurs in  $O(n^2)$  time, but in fact none of the vertices is added again to the  $UCH(X)$  after removal. It follows that the asymptotic estimate for the main cycle run is  $O(n)$  in total and it is the ordering in the initialization dominating with its  $O(n \log n)$  time. Clearly the linear  $O(n)$  memory is sufficient and so the optimal solution is achieved again for the convex hull problem.

**13.3.3. The divide and conquer paradigm.** Another very standard idea is to divide the entire problem into pieces, apply recursively the same procedure on them, and merge the partial results together. These are the two phases of the *divide and conquer* approach. This paradigm is common in many areas, cf. 13.I.10.



With convex hulls, adopt the gift wrapping approach for the conquer phase. The idea is to split recursively the task producing disjoint “left  $CH(X)$ ” and “right  $CH(X)$ ” and to

**13.I.25.** How many sequences of  $\{0, 1, 2, 3\}$  of length  $n$  have at least 2 zeroes?

**Solution.** Exponential generating function for 0- entry is

$$b_0(x) = \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = e^x - 1 - x$$

. Exponential generating function for other entries is  $e^x$  as there are no restrictions. Hence, the exponential generating function in question is  $f(x) = e^{4x} - e^{3x} - x e^{3x}$ . The answer to the question is  $n!$  times the coefficient at  $x^n$  in  $f(x)$ , which is

$$n! \left( \frac{4^n}{n!} - \frac{3^n}{n!} - n \frac{3^{n-1}}{n!} \right) = 4^n - 3^n - n3^{n-1}.$$

□

merge them by finding the upper and lower “tangent edge” of those two parts.

DIVIDE AND CONQUER CONVEX HULL

*Input:* Set of points  $X = \{v_1, \dots, v_n\}$  in the plane, ordered lexicographically.

*Output:* The directed path  $CH(X)$ .

(1) *Divide.* If  $n \leq 3$ , return the  $CH(X)$ . Otherwise, split  $X = X_1 \cup X_2$  into two subsets of (roughly) same sizes, respecting the order (i.e. all vertices in  $X_1$  smaller than those in  $X_2$ ).

(2) *Merge.*

- Start with the edge joining the largest point in  $CH(X_1)$  and the smallest in  $CH(X_2)$ , and iteratively balance it to the lower tangent segment  $e_l$  to  $CH(X_1)$  and  $CH(X_2)$ .
- Proceed similarly to get the upper tangent  $e_u$ .
- Merge the relevant parts of the  $CH(X_1)$  and  $CH(X_2)$  with the help of  $e_l$  and  $e_u$ .

Perhaps the merge step requires some more explanation. The situation is illustrated in the diagram. For the upper tangent, first fix the right-hand vertex of the initial edge joining the two convex polygons. Then find the tangent to the left polygon from this vertex. Then fix the head of the moving edge and find the right hand touch point of the potential tangent. After a finite number of exchanges like this, the edge stabilizes. This is the upper tangent edge  $e_u$ . Observe that during the balancing, we move only clockwise on the right-hand polygon and counter-clockwise on the other one. Notice also that it is the smart divide strategy which prevents any of the points of the input  $X_1$  to appear inside of the  $CH(X_2)$  and vice versa.

Again the analysis of the algorithm is easy. The typical merge time is asymptotically linear and then the recursive call yields  $O(\log n)$  runs of the procedures. The total estimated time is again  $O(n \log n)$ . Notice, there is no initialization in the procedure itself. Just assume that the points in  $X$  are already ordered. Hence another  $O(n \log n)$  time must be added to prepare for the very first call. The memory necessary for running the algorithm is estimated by  $O(n)$  if the recursions are implemented properly.

**13.3.4. The incremental paradigm.** This approach consists in taking the input objects one by one and consecutively build the required resulting structure. This is particularly useful if the application does not allow to have all the points available in the beginning. Imagine *incrementally* building the convex hull of shots into the target, as they occur.

Another good use is in the *randomized algorithms*, where all the data is known in the beginning, but treated in a random order. Typically the expected time for running is then very good, while there might be much less effective, but improbable, worst time runs.



The former case is easy to illustrate on the convex hull problem. In each step employ the merge step of the very degenerate version of the divide and conquer algorithm, merging  $CH(X_1)$  of the previously known points  $X_1$ , while  $X_2 = CH(X_2) = \{v_k\}$  is just the new point. But an extra step is needed to check whether  $v_k$  is inside of  $CH(X_1)$  or not. If it is, then skip the new point and wait for the next one. If not, then merge. The worst time of this algorithm is  $O(n^2)$ , but as with the gift wrapping method, it depends on the actual size of the output as well as on the quality of the algorithm checking whether  $v_k$  is inside of the  $CH(X_1)$ .

We illustrate the second case with a more elaborate convex hull algorithm. The main idea is to keep track of the position of all points in  $X$  with respect to the convex hull  $CH(X_k)$  of the first  $k$  of them (in the fixed order chosen at the beginning).

With this goal in mind, keep the dynamical structure of a bipartite graph  $G$  whose first group of vertices consists of those points which have not been processed yet, while the other group contains all the faces of the current convex polygon  $S = CH(X_k)$  (call them faces, not to be confused with the edges in the graph in  $G$ ). Remember the faces in  $S$  are oriented. Such a face  $e$  is in conflict with the point  $v$  if the face is “visible” from  $v$ , i.e.  $v$  is in the right-hand halfplane determined by  $e$ . Keep all points joined to each of their faces in conflict in the bipartite graph. Call  $G$  the *graph of conflicts*. The algorithm can now be formulated:

#### RANDOMIZED INCREMENTAL CONVEX HULL ALGORITHM

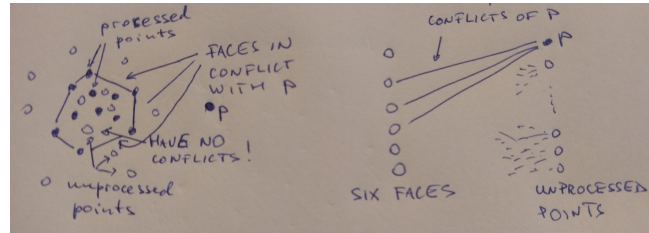
*Input:* A set  $X = \{v_1, \dots, v_n\}$ ,  $n > 3$ , of at least four points in the plane.

*Output:* The edge list  $R$  of the convex hull  $CH(X)$ .

- (1) *Initialization.* Fix a random order on  $X$ . Choose the first three points as  $X_0$ , create the list of conflicts for the edge list  $R = CH(X_0)$  (i.e. state which of the three faces are seen from which points) and remove the three points from  $X$ .
- (2) *Main cycle.* Repeat until the list  $X$  is empty:
  - choose the first point  $v \in X$ ;
  - if there are some conflicts of  $v$  in  $G$ , then
    - remove all the faces in conflict with  $v$  from both  $R$  and  $G$ ,
    - find the two new faces (the upper and lower tangents from the new point  $v$  to the existing  $CH(X)$  – they are easily found by taking care of the “vertices without missing edges”),
    - add the two new faces to both  $R$  and  $G$  and find all their conflicts;
  - remove the point  $v$  from the list  $X$ , and from the graph  $G$ .

The complete analysis of this algorithm is omitted. Notice that finding the newly emerging conflicts is easy since it is only necessary to check the potential conflicts of points

which are in conflict with the faces incident to those two vertices in  $G$  before the update, to which the two new faces are attached.



It can be proved that the expected time for this algorithm is  $O(n \log n)$ , while the worst time is  $O(n^2)$ . The complete framework for the analysis of randomized algorithms is nicely explained in the book mentioned in the very beginning of this part, see page 1188.

**13.3.5. Convex hull in 3D.** Many applications need convex hulls of finite sets of points in higher dimensions, in particular in 3D. There are several ways of adjusting 2D algorithms to their 3D versions.



First, it needs to be stated what the right structure is for the  $CH(X)$ . As seen in 13.1.21, the convex polyhedra in  $\mathbb{R}^3$  can be perfectly described by planar graphs. In order to modify the algorithms into 3D, some good encoding for them is needed. We want to find all vertices with edges or faces which are incident or neighbouring in time proportional to the output.

This is nicely achieved by the *double connected edge lists*.

DOUBLE CONNECTED EDGE LIST – DCEL

Let  $G = (V, E, S)$  be a planar graph. The *double connected edge list* is the list  $E$  such that each edge is represented by two oriented twin-edges  $e, \bar{e}$ , equipped with the pointers

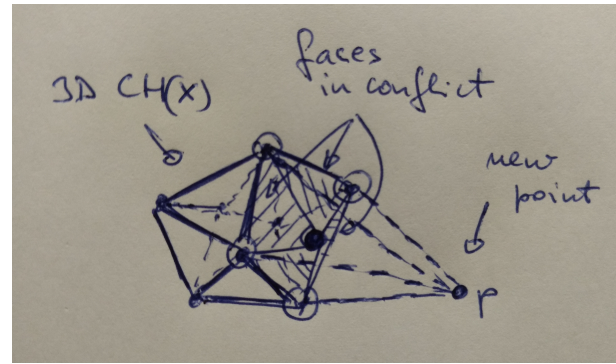
- $V_t, V_h$  to tail and head of  $e$
- $F$  to the incident face (the left one with respect to the directed edge  $e$ ) and the pointer to the twin  $\bar{e}$
- $P$  to the edge following along the face  $F$  (in the counter clockwise direction).

At the same time, keep the list of vertices and the list of faces, always with just one pointer towards some of the incident edges.

Clearly, we can remove or add faces, edges, vertices, or find all kind of neighbors in time proportional to the output (notice the use of the twins and think about the details!).

Next, look at the 2D incremental algorithm above and try to imagine what needs changing there to get it to work in 3D. First, we have to deal with the 2D faces  $S$  of the DCEL of the convex hull and instead of their boundary vertices, deal with boundary edges. Again, all the faces with conflicts with the just processed point have to be removed (see the picture). This leads to a directed cycle of edges the pointers  $F$  missing (call them “unsaturated edges”). Finally, instead of adding

two new faces in 2D, add the tangent cone of faces joining the point  $v$  with the unsaturated edges. Of course the graph of conflicts must also be updated.



#### RANDOMIZED INCREMENTAL 3D CONVEX HULL ALGORITHM

*Input:* A set  $X = \{v_1, \dots, v_n\}$ ,  $n \geq 4$ , of at least five points in the space  $\mathbb{R}^3$ .

*Output:* The DCEL  $R$  for the convex hull  $CH(X)$ .

- (1) *Initialization.* Fix a random order on  $X$ . Choose the first four points as  $X_0$ , create the list of conflicts  $G$  and the DCEL  $R$  for  $CH(X_0)$  (i.e. tell which of the four faces are seen from which points) and remove the four points from  $X$ .
- (2) *Main cycle.* Until the list  $X$  is empty, repeat:
  - take the first point  $v \in X$ ;
  - if there are some conflicts of  $v$  in  $G$ , then
    - remove all the faces of  $R$  in conflict with  $v$  (from both  $R$  and  $G$ ), and take care of the (oriented) edges  $e$  in  $R$  left without their incident faces,
    - build the “tangent cone” from the new point  $v$  to the current  $R$  by connecting  $v$  to the latter “unsaturated” edges,
    - add the new faces to both  $R$  and  $G$  and find all their conflicts (again, note that the check for new conflicts can be restricted to the points which were in conflict with the faces incident to those edges where the cone has been attached to the previous  $R$ ).
  - remove the point  $v$  from the list  $X$  and from the graph  $G$ .

A detailed analysis is omitted. As with the 2D case, the expected running time for this algorithm is  $O(n \log n)$ . By the very well adapted DCEL data structure for the convex hull, it is a very good algorithm.

The divide and conquer algorithm from the 2D case can be easily adapted, too. Skipping details, the initial ordering of the input points lexicographically allows to recursively call the same procedure producing two DCELs of disjoint convex polytopes. This allows us to apply a more sophisticated “gift wrapping” approach when merging the results. A sort of “tubular collar” wrapping of the two polytopes to create their



convex hull is desired. Imagine rotating the hyperplanes similarly as with the lines in 2D in order to get the first edge in the tubular set of faces to be added. Then the first plane containing one of the missing new faces is obtained. Continue breaking the plane along the new edges, until both directed cycles along of which the collar is attached to the previous two polytopes are closed. All that is done by bending the planes by the smallest possible angle in each step and checking to arrive at the right position. Of course, the DCEL structure is essential to update all the data properly in time proportional to the size of the changes.

With reasonably smart implementation, this algorithm achieves the optimal  $O(n \log n)$  running time.

Both of the latter algorithms can be generalized to all higher dimensions, too.

**13.3.6. Voronoi diagrams.** The next stop is at one of



the most popular and useful planar divisions (and searching in them). For a finite set of points  $X = \{v_1, \dots, v_n\}$  in the plane  $\mathbb{R}^2$ , write  $VR(v_i)$  for the set of points in  $\mathbb{R}^2$  sharing  $v_i \in X$  as their uniquely given closest point in  $X$ . Define:

VORONOI DIAGRAM

For a given set of points  $X = \{v_1, \dots, v_n\}$  (not all colinear), the *Voronoi regions* are

$$VR(v_i) = \{x \in \mathbb{R}^2; \|x - v_i\| < \|x - v_k\|, \text{ for all } v_k \in X, k \neq i\}.$$

This is an intersection of  $n - 1$  open half-planes bounded by lines, so it is an open convex set. Its boundary is a convex polygon.

The *Voronoi diagram*  $VD(X)$  is the planar graph whose faces are the open regions  $VR(v_i)$ , while the boundaries of  $VR(v_i)$  yield the edges and vertices.

4

Care is needed about collinearity since if all the points  $v_i$  are on same line in  $\mathbb{R}^2$ , then their Voronoi regions are strips in the plane bounded by parallel lines. Under all other circumstances, the planar graph  $VD(X)$  from the latter definition is well defined and connected.

By definition, the vertices  $p$  of  $VD(X)$  are the points in the plane, such that at least three points  $v, w, u \in X$  are the same distance from  $p$  and no more points of  $X$  are inside the circle through  $v, w, u$ . If there are no more points of  $X$  on the latter circle, then the degree of this vertex  $p$  is 3.

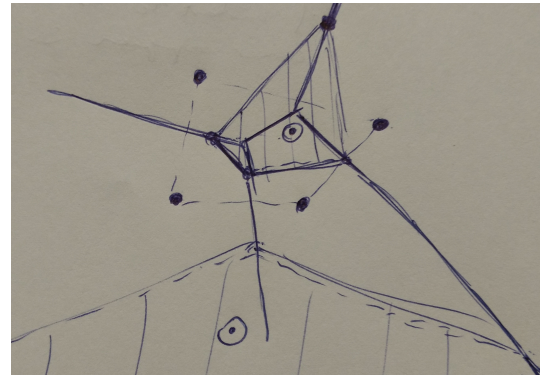
The most degenerate situation occurs if all the points of  $X$  are on one circle. Then, obviously, the Voronoi regions are delimited by two half lines, all emanating from the center of the circle and cutting the angles properly. The construction of the  $VD(X)$  is then equivalent to ordering of the points by angles. At least  $O(n \log n)$  time is needed for the worst case estimate in any algorithm building the Voronoi diagrams.

Some of the Voronoi regions are unbounded, others bounded. If just two points  $v$  and  $w$  are considered, then the axis of the segment  $vw$  is the boundary for the regions

$VR(v)$  and  $VR(w)$ . In particular, the region  $VR(v)$  must be bounded for each  $v$  in the interior of the convex hull of  $X$ . On the contrary, consider an edge in the  $CH(X)$  with incident vertices  $v$  and  $w$ , and the “outer” half-axis of the segment  $vw$ . If one considers any interior point  $u$  in the  $CH(X)$ , then it is in the other halfplane with respect to the segment  $vw$ . Sooner or later the points in the latter half-axis are closer to  $v$  and  $w$  than to  $u$ . It follows that both  $VR(v)$  and  $VR(w)$  are unbounded. Summarizing:

**Lemma.** *Each Voronoi region  $VR(v)$  of the Voronoi diagram  $VD(X)$  is an open convex polygon region. It is unbounded if and only if  $v$  belongs to the convex hull of  $X$ .*

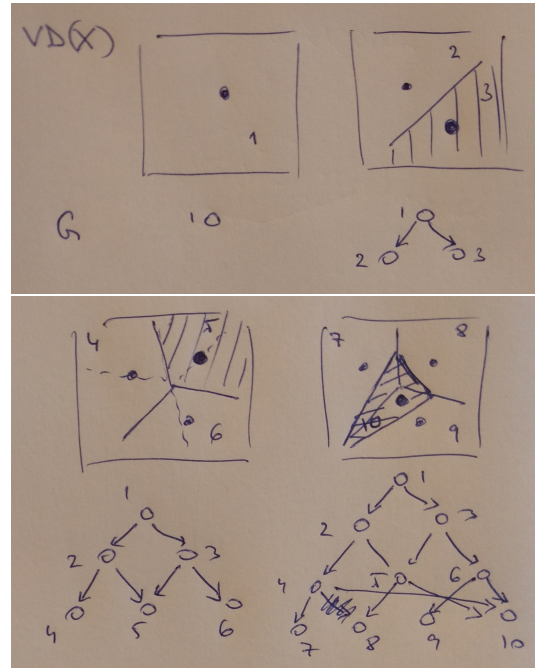
**13.3.7. An incremental algorithm.** Each Voronoi diagram represents a planar division and adding a new point  $p$  as a vertex, it is quite obvious how to update the diagram. First assume, we know which  $VR(v)$  the point  $p$  hits. Then choose the center  $v$ , split the region  $VR(v)$  by the relevant part of the axis of the segment  $pv$ . Add this new edge  $e$  into the updated  $VD(X)$ , simultaneously creating the two new faces and removing the  $VR(v)$  one. The new edge  $e$  hits the boundary of the current  $VR(v)$  in either two points or one point (if the new edge is unbounded). These hits show what is the next region of the updated diagram to be split. “Walk” further with the new hit at the boundary revealing the next center of region playing the role of  $v$  above. Ultimately this walk consecutively splits the visited old regions and creates the new directed cycle of edges bounding the new region, or it has an unbounded path of boundary edges, if the new region is unbounded. See the diagram for an illustration.



If the new point  $p$  is on the boundary, i.e. hitting one of the edges or vertices in  $VD(X)$ , then the same algorithm works. Just start with one of the incident regions.

So far this looks easy, but how does one find the relevant region hit by the new point? An efficient structure to search for it on the run of the algorithm is desired. Build an acyclic directed graph  $G$  for that purpose. The vertices in  $G$  are all the temporary Voronoi regions as they were created in the individual incremental steps. Whenever a region is split by the above procedure, new leaves in  $G$  are created. Draw edges towards these leaves from all the earlier regions which have some nontrivial overlap. Of course, care must be taken how

the old regions are overlap with the new ones, but this is not difficult. We illustrate the procedure on the diagram, updating from one point to four points.



INCREMENTAL VORONOI DIAGRAM

*Input:* The set of points  $X = \{v_1, \dots, v_n\}$  in the plane, not all collinear.

*Output:* The DCEL of  $VD(X)$  and the search graph  $G$ .

- (1) *Initialization.* Consider the first two points  $X_0 = \{v_1, v_2\}$  and create the DCEL for  $VD(X_0)$  with two regions. Create the acyclic directed graph  $G$  (just root and two leaves).
- (2) *Main cycle.* Repeat until there are no new points  $z \in X$ :
  - localize the  $VR(v)$  hit by  $z$  (by the search in  $G$ )
  - perform the path walk finding the boundary of the new region  $VR(z)$  in  $VD(X)$
  - update the DCEL for  $VD(X)$  and the acyclic directed search graph  $G$ .

This algorithm is easy to implement. It produces directly a search structure for finding the proper Voronoi regions of given points. Unfortunately, it is very far from optimal in both aspects - the worst running time is  $O(n^2)$ , and the worst depth of the search acyclic graph is  $O(n)$ . If this is treated as a randomized incremental algorithm, the expected values is better, but not optimal either. Below is a useful modification via triangulations.

**13.3.8. Delaunay triangulation.** One remarkable feature



of the Voronoi diagrams should not remain unnoticed. Right after the definition of the Voronoi diagram, an important fact was mentioned. The vertices of the planar graph  $VD(X)$  are centers of circles containing at least three points

of  $X$ , and no other points of  $X$  are inside of the circle. The dual graph to  $VD(X)$  (see 13.1.22 for the definition) can be realized by taking the vertices  $V = X$  and the edges indicate the neighboring faces. Then this is again a tessellation of the plane. Actually, the convex hull of  $X$  is tessellated into convex regions, and one unbounded region remains. It is called the Delaunay tessellation  $DT(X)$ . In the generic case, the degrees of all vertices in  $VD(X)$  are 3 (i.e. no four points of  $X$  lie on one circle). This is the *Delaunay triangulation* of the convex hull of  $X$ .<sup>13</sup> Notice that it is easy to turn any Delaunay tessellation into a triangulation by adding the necessary edges to triangulate the convex regions with more edges. Any of these refined tessellations is called the Delaunay triangulation associated to the  $VD(X)$ .

In general, a planar graph  $T$  is called a *triangulation* of its vertices  $X \subset \mathbb{R}^2$ ,  $|X| = n$ , if all its bounded faces have just 3 vertices. It is easy to see that each triangulation  $T$  has  $\tau = 2n - 2 - k$  triangles and  $\nu = 3n - 3 - k$  edges, where  $k$  is the number of vertices in the boundary of the unbounded face. Indeed, by the Euler formula (13.1.20)  $n - \nu + \tau + 1 = 2$  (there is an unbounded face on top of all the triangles). Now, every triangle has 3 edges, while there are  $k$  edges around the unbounded face. It follows that  $3\tau + k = 2\nu$ . It remains to solve the two linear equations for  $\tau$  and  $\nu$ .

The triangulations are extremely useful in numerical mathematics and in computer graphics as the typical background mesh for processing of approximate values of functions. Of course, there are many triangulations on a given set and one of the qualitative requests is to aim at triangles as close to the equilateral triangles as possible. This could be phrased as the goal to maximize the minimal angles inside the triangles.

A practical way to do this is to write the *angle vector of the triangulation*

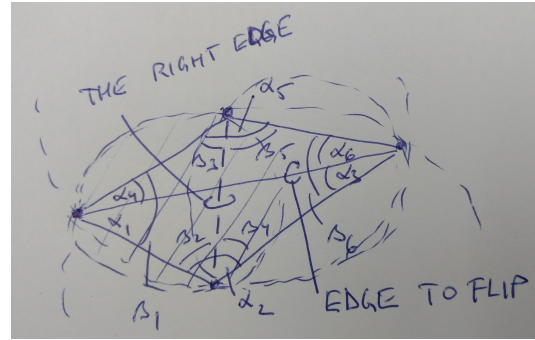
$$\mathcal{A}(T) = (\gamma_1, \gamma_2, \dots, \gamma_{3\tau}),$$

where  $\gamma_j$  are the angles of all the triangles in  $T$  sorted by their value,  $\gamma_j \leq \gamma_k$  for all  $j < k$ .

A triangulation  $T$  on  $X$  is said to be *angle optimal*, if  $\mathcal{A}(T) \geq \mathcal{A}(T')$  for all triangulations  $T'$  on the same set of vertices  $X$ , in the lexicographic ordering. In particular, an angle optimal triangulation achieves the maximum over the minimal angles of the triangles.

Surprisingly, there is a very simple (though not very effective) procedure to produce (one of) the angle optimal triangulations. Consider any two adjacent triangles and check the six angle sequences of their interior angles. If the current position of the diagonal edge provides the worse sequence, flip it. See the diagram.

<sup>13</sup>Although the name sounds French, Boris Nikolaevich Delone (1890 - 1990) was a Russian mathematician using the French transcription of his name in his publications. His name is associated with the triangulation because of his important work on this from 1934.



The flip is necessary if and only if one of the vertices outside the diagonal is inside the circle drawn through the remaining three vertices.

Since each such flipping of an edge inside of a triangulation  $T$  definitively increases the angle vector, the following algorithm must stop and achieve an angle optimal triangulation:

EDGE FLIPPING DELAUNAY

*Input:* Any triangulation  $\tilde{T}$  of points  $X$  in the plane.  
*Output:* An angle minimal triangulation  $T$  of the same set  $X$ .

(1) *Main cycle.* Repeat until there are no edges to flip:

- Find an edge which should be flipped and flip it.

**Theorem.** A triangulation  $T$  on a set of points  $X$  in the plane  $\mathbb{R}^2$  is angle optimal if and only if it is a Delaunay triangulation associated to the Voronoi diagram  $VD(X)$ .

**PROOF.** Consider any Delaunay triangulation  $T$  associated to  $VD(X)$  and one of the vertices  $p$  of  $VD(X)$ . If the four vertices of two neighboring triangles are not all on the same circle, then by the very definition of  $VD(X)$ , the two circles in question do not contain the remaining points and thus there is no need for any flips. Further, let  $v_1, \dots, v_k$  be all the points of  $X$  lying on the circle determining  $p$ . Fix an edge with two neighbouring endpoints on a circle. All triangles with the third vertex on the circle above the edge share the same angle. A simple check now verifies that different ways of triangulating the same region of  $VD(X)$  with more than 3 boundary edges lead always to the same angle vector.

In particular, there are no flips at all necessary in the above algorithm if one starts with the Delaunay triangulation  $T$ . Hence the angle optimal triangulation is arrived at.

In order to prove the other implication, recall the comments on the diagram above. All triangles in the angle optimal triangulations  $T$  have the following two properties: (1) the circle drawn through their three vertices do not include any other point in its interior: (2) the circle having any of their edges as diameter does not have any other point in its interior.

Consider the dual graph  $G$  of  $T$  and consider its realization in the plane by drawing the vertices as the centers of the circles drawn through the vertices of individual triangles,

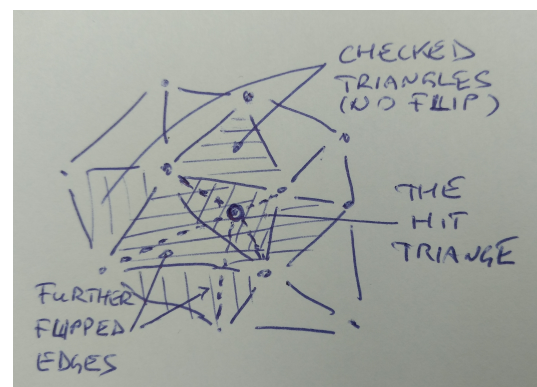
while the edges are the segments joining them. If there are not more than 3 points on any of those circles, then  $G = VD(X)$  is obtained. In the degenerate situations, all the triangles sharing the circle produce the same vertex in the plane and some of the relevant edges degenerate completely. Identify those collapsing elements in  $G$ , to get the right  $VD(X)$ .  $\square$

**13.3.9. Incremental Delaunay.** We return to the general idea for the Voronoi diagram, namely to design an algorithm which constructs both  $VD(X)$  and  $DT(X)$  and which behaves very well in its randomized implementation.



The idea is straightforward. Use the incremental approach as with the Voronoi diagrams for refining the consecutive Delaunay triangulations, employing the flipping of edges method.

By looking at the diagram, the Voronoi algorithm is easily modified. Care must be taken of three different cases for the new points – hitting the unbounded face, hitting one of the internal triangles, hitting one of the edges.



## INCREMENTAL DELAUNAY TRIANGULATION

*Input:* The set of points  $X = \{v_1, \dots, v_n\}$  in the plane, not all collinear.

*Output:* The DCEL of both  $DT(X)$  and the search graph  $G$  for this triangulation.

- (1) *Initialization.* Consider the first three points  $X_0 = \{v_1, v_2, v_3\}$ . Create the DCEL for  $DT(X_0)$  with two regions, and create the  $CH(X_0)$  (the connected edge list). Create the acyclic directed graph  $G$  (just root and two leaves).
- (2) *Main cycle.* Repeat until there are no new points  $z \in X$ :
  - Localize the face  $\Delta$  in  $DT(X_k)$  hit by  $z$  (by the search in  $G$ )
  - if  $z$  is in the unbounded face, then
    - add the new triangles  $\Delta_1, \dots, \Delta_\ell$  to  $DT(X_k)$  by joining  $z$  to visible edges in  $CH(X_k)$
    - update the  $CH(X)$ .
  - if  $z$  hits a (bounded) triangle  $\Delta$ , then split it into the three new triangles  $\Delta_1, \Delta_2, \Delta_3$ .
  - if  $z$  hits an edge  $e$ , then split the adjacent bounded triangles into  $\Delta_1, \dots, \Delta_4$  (only two, if an edge in  $CH(X_k)$  is hit).
  - Create a queue  $Q$  of not yet checked modified triangles and repeat as long  $Q$  is not empty:
    - take the first  $\Delta$  from the queue  $Q$ , look for its neighbour not in  $Q$  and not yet checked, and flip the edge if necessary;
    - if an edge is flipped, put the newly modified triangles into  $Q$

Detailed analysis of the algorithm is omitted. It is almost obvious that the algorithm is correct. It is only necessary to prove that the proposed version of the flip edge algorithm update ensures that after each addition of the new point  $z$  in  $k$ th step, the correct Delaunay triangulation of  $X_k$  arises. Once an edge is flipped, then it is not necessary to consider it any time later.

Finally, if the Voronoi diagram is needed instead, it can be obtained from the  $DT(X)$  in linear time. Obviously the search structures can be used directly.

Surprisingly enough, it turns out that the expected number of total flips necessary over all the run is of size  $O(n \log n)$ . Hence the algorithm achieves the perfect expected time  $O(n \log n)$ . Detailed analysis of this beautiful example of results in computational geometry can be found in the section 9.4. of the book by Berg et al., cf. the link on page 1188.

**13.3.10. The beach line Voronoi.** The Voronoi algorithm provides a perfect example for the sweep line paradigm, where the extra structure to keep track of the events has to be quite smart.<sup>14</sup>

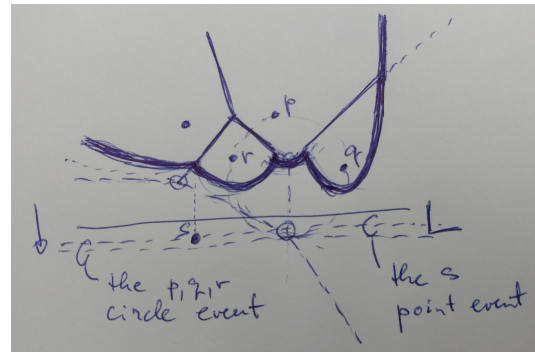


<sup>14</sup>This is mostly called the *Fortune Voronoi algorithm*. Not because this is such a lucky construction, but rather because the algorithm was published by Steven Fortune of the Bell Laboratories in 1986

Imagine the horizontal line  $L$  (parallel to the  $x$  axis) flows in the top-down direction and meets the points in  $X = \{v_1, \dots, v_n\}$ . Of course  $VD(X)$  including all points above the current position of  $L$  cannot be drawn, since it depends also on the points below the line. It is better to look at the part  $R_L$  in the plane,

$$R_L = \{p \in \mathbb{R}^2; \text{dist}(p, L) \geq \text{dist}(p, v_i), v_i \in X \text{ above } L\}.$$

This is exactly the part of  $\mathbb{R}^2$  which can be tessellated into the Voronoi diagram with the information collected at the current position of  $L$ . Obviously,  $R_L$  is bounded by a continuous curve consisting of parts of parabolas, since for one point  $v_i$  this is the case, and the intersection of the conditions is relevant.



Call the boundary of  $R_L$  the *beach line*  $B_L$ . The vertices on  $B_L$  draw the  $VD(X)$  when  $L$  is moving. Since the Voronoi diagram consists of lines, we do not even compute the parabolas and take care of the arrangements of the still active parts of parabolas in the beachline, as determined by the individual points. New parts of the beachline arise when the line  $L$  meets one of the points. Add all the points to an ordered list in the obvious lexicographic order. Call them the *point events*. The active arc in the beachline disappears if the line  $L$  meets the bottom of the circle drawn through three points determining a vertex in the Voronoi diagram. Such an event is called a *circle event*.

Both types of the events are illustrated in the diagram above. There is a striking difference between them.

The point events always initiate a new arc and start “drawing” two edges of the Voronoi diagram. They initiate previously unknown circle events.

The circle events might disappear without creating a genuine vertex in  $VD(X)$ . Look at the diagram at the  $s$  point event. The new  $s, r, q$  circle event is encountered there. But this would not create a vertex in the diagram if there was the next point  $u$  somewhere close enough to the indicated vertex. One could find it out as soon as such a point event  $u$  is met. Such “ineffectively disappearing” circle events are called *false alarms*. On the contrary the  $p, q, r$  circle event shown in the diagram gives rise to the indicated vertex.

Summarizing, the emerged circle events must be inserted properly into the ordered queue of events and handled properly at each of the point events.



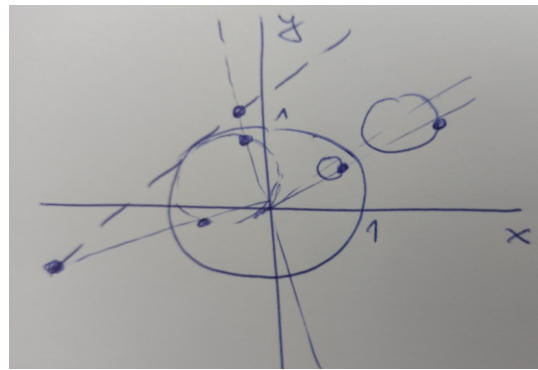
Further details are not considered here. When implemented properly, this algorithm runs in the optimal  $O(n \log n)$  time and  $O(n)$  storage. See again the above mentioned book by Berg et al. (section 7) for details.

**13.3.11. Geometric transformations.** Various geometric transformations of the basic ambient space can often help to transform one problem into another one. This is illustrated in a beautiful construction relating the convex hulls and the Voronoi diagrams.



Of course transformations which behave well on lines and planes and preserve incidences should be well thought of. The affine and projective transformations behave well in this respect as in the fourth chapter. Introduce a more interesting one – the *spherical inversion*. In the plane  $\mathbb{R}^2$ , consider the unit circle  $x^2 + y^2 = 1$ . For arbitrary  $v = (x, y) \neq (0, 0)$  define

$$\varphi(v) = \frac{1}{\|v\|^2} v = \frac{1}{x^2 + y^2} (x, y)$$



Clearly  $\varphi$  is a bijection on  $\mathbb{R}^2 \setminus \{(0, 0)\}$ . The geometric meaning of such a transform is clear from the formula, see the diagram. "The general" point  $v$  is sent to a point on the same line through the origin, but with reciprocal size. The unit sphere is the set of all fixed points.

The same principle works for all dimensions, so we may equally well (and more interestingly) consider  $v \in \mathbb{R}^3$  in the sequel.

Next follows the crucial property of  $\varphi$ .

**Lemma.** *The mapping  $\varphi$  maps the spheres and planes in  $\mathbb{R}^3$  onto spheres and planes. The image of a sphere is a plane if and only if the sphere contains the origin.*

**PROOF.** Consider a sphere  $C$  with the center  $c$  and radius  $r$ . The equation for its general points  $p$  reads

$$\|p - c\|^2 = r^2.$$

By drawing a few images as in the diagram above, it is easily guessed that the images will be a circle with the center  $s = \frac{1}{\|c\|^2 - r^2} c$  (i.e. again on the same line through the origin). Now consider  $q = \varphi(p)$  and compute (using just

$2p \cdot c = \|p\|^2 + \|c\|^2 - r^2$  from the latter equation)

$$\begin{aligned} \|q - s\|^2 &= \left\| \frac{p}{\|p\|^2} - \frac{c}{\|c\|^2 - r^2} \right\|^2 \\ &= \frac{1}{\|p\|^2} + \frac{1}{(\|c\|^2 - r^2)^2} \|c\|^2 - 2 \frac{p \cdot c}{(\|c\|^2 - r^2)\|p\|^2} \\ &= \frac{1}{(\|c\|^2 - r^2)^2} \|c\|^2 - \frac{1}{\|c\|^2 - r^2} \\ &= \left( \frac{r}{\|c\|^2 - r^2} \right)^2. \end{aligned}$$

The latter computation assumes  $\|c\| \neq r$ . Fix the center  $c \neq 0$  and consider diameters  $r$  approaching  $\|c\|$  from below or above. Then the images are circles with the centers  $s$  approaching one or the other infinite points of the line  $\text{span}\{c\}$  and fast growing radii. In the limit position, the plane is obtained as requested. (Check this asymptotic computation directly yourself, if any doubts.)  $\square$

The continuity of  $\varphi$  has got important consequences. Consider a general plane  $\mu$  (not containing the origin). The inversion  $\varphi$  maps one of the half-spaces determined by  $\mu$  to the interior of the image sphere. The other half-space maps to the unbounded complement of the sphere. The latter is of course the half-space containing the origin.

The efficient link between the Voronoi diagrams and convex hulls can now be explained. Assume a set of points  $X = \{v_1, \dots, v_n\}$  in the plane is given. View them as the points in the plane  $z = 1$  in  $\mathbb{R}^3$ , i.e. add the same coordinates  $z = 1$  to all the points  $(x, y)$  in  $X$ . For simplicity, assume that no three of them are collinear and no four of them lie on the same circle.

The spherical inversion  $\varphi$  maps the entire plane  $z = 1$  to the sphere  $S$  with center  $c = (0, 0, 1/2)$  and radius  $1/2$ . Write  $w_1, \dots, w_n$  for the images  $w_i = \varphi(v_i)$ .

Now, consider  $CH(Y)$  for the set of the images  $Y = \{w_1, \dots, w_n\}$ . This is a convex polytope with all vertices on the sphere  $S$ . All its faces represent planes not containing the origin (this is due to the assumption that no three points of  $X$  are collinear).

Split the faces of  $CH(Y)$  into those “visible” from the origin and the “invisible” ones. In the latter case, all points of  $Y$  are on the same side of plane  $\mu$  generated by the face as the origin. This implies that all the other points are outside of the image sphere  $S_\mu = \varphi(\mu)$ . In particular, there are no points of  $X$  inside of the intersection of  $S_\mu$  with the plane  $z = 1$ . This is the defining condition for obtaining one of the vertices of the Voronoi diagram. Since the map  $\varphi$  preserves incidences, the entire DCEL for  $VD(X)$  is easily reconstructed from the DCEL of  $CH(Y)$  and vice versa.

This resembles the construction of the dual graph, i.e. the Delaunay triangulation  $DT(X)$  from the Voronoi diagram, with further geometric transformation in the back.

Last, but not least, the faces of  $CH(Y)$  visible from the origin are worth mentioning too. For the same reason as above, all the points of  $Y$  appear on the other side from

the origin and so, all the points in  $X$  are inside the image sphere. This means the diagram of furthest points instead of the Voronoi diagram of the closest ones is obtained.

This is a very useful tool in several areas of mathematics, see some of the exercises (??) for further illustration.

**4. Remarks on more advanced combinatorial calculations**

**13.4.1. Generating functions.** The worlds of discrete and continuous mathematics meet all the time.



There are already many instances of useful interactions. With some slight exaggeration, we can claim that all results in

analysis were achieved by an appropriate reduction of the continuous tasks to some combinatorial problem (for instance, integration of rational functions is reduced to partial fraction decomposition, solution of analytic differential equations possibly boils down to some recurrences, etc.). In the opposite direction, we demonstrate how handy continuous methods can be in purely combinatorial problems.

We begin with a simple combinatorial question: *There are four 1-crown coins, five 2-crown coins, and three 5-crown coins at our disposal. Suppose we want to buy a bottle of coke which costs 22 crowns. In how many ways can we pay the exact amount of money with the given coins?*

We are looking for integers  $i, j, k$  such that  $i+j+k = 22$  and

$$i \in \{0, 1, 2, 3, 4\}, j \in \{0, 2, 4, 6, 8, 10\}, k \in \{0, 5, 10, 15\}.$$

Consider the product of polynomials (over the real numbers, for instance)

$$(x^0 + x^1 + x^2 + x^3 + x^4)(x^0 + x^2 + x^4 + x^6 + x^8 + x^{10})(x^0 + x^5 + x^{10} + x^{15}).$$

It should be clear that the number of solutions equals the coefficient at  $x^{22}$  in the resulting polynomial. This corresponds to the four possibilities of choosing the values  $i, j, k$ :  $3 \cdot 5 + 3 \cdot 2 + 1 \cdot 1$ ,  $3 \cdot 5 + 2 \cdot 2 + 3 \cdot 1$ ,  $2 \cdot 5 + 5 \cdot 2 + 2 \cdot 1$ , and  $2 \cdot 5 + 4 \cdot 2 + 4 \cdot 1$ .

This simple example deserves more attention.

The coefficients of the particular polynomials represent sequences of numbers, referring to how many times we can achieve the given value with one type of coins only. Work with an infinite sequence to avoid a prior bound on how many available values there can be. Encode the possibilities in infinite sequences

$$(1, 1, 1, 1, 1, 0, 0, \dots) \quad \text{1-crowns}$$

$$(1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, \dots) \quad \text{2-crowns}$$

$$(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, \dots) \quad \text{5-crowns.}$$

Each such sequence with only finitely many non-zero terms can be assigned a polynomial. The solution of the problem is given by the product of these polynomials, as noted before.

This is an instance of a general procedure for handling sequences effectively.

## GENERATING FUNCTION OF A SEQUENCE

**Definition.** An (ordinary) *generating function* for an infinite sequence  $a = (a_0, a_1, a_2, \dots)$  is a (formal) power series

$$a(x) = a_0 + a_1x + a_2x^2 + \dots = \sum_{i=0}^{\infty} a_i x^i.$$

The values  $a_i$  are considered in some fixed field  $\mathbb{K}$ , normally the rational numbers, real numbers, or complex numbers.

In practice, there are several standard ways for defining and using generating functions:

- to find an *explicit formula* for the  $n$ -th term of a sequence;
- to derive *new recurrent relations* between values (although generating functions are often based on recurrent formulae themselves);
- for *calculation of means* or other statistical dependencies (for instance, the average time complexity of an algorithm);
- to prove *miscellaneous combinatorial identities*;
- to find an *approximate formula* or the *asymptotic behaviour* when the exact formula is too hard to get.

We shall see examples of some of these.

### 13.4.2. Operations with generating functions.



Several basic operations with sequences correspond to simple operations over power series (which can be easily proved by performing the relevant operation with the power series):

- Component wise, the sum  $(a_i + b_i)$  of the sequences corresponds to the sum  $a(x) + b(x)$  of the generating functions.
- Multiplication  $(\alpha \cdot a_i)$  of all terms by a given scalar  $\alpha$  corresponds to the same multiplication  $\alpha \cdot a(x)$  of the generating function.
- Multiplication of the generating function  $a(x)$  by a monomial  $x^k$  corresponds to shifting the sequence  $k$  places to the right and filling the first  $k$  places with zeros.
- In order to shift the sequence  $k$  places to the left (i.e. omit the first  $k$  terms), subtract the polynomial  $b_k(x)$  corresponding to the sequence  $(a_0, \dots, a_{k-1}, 0, \dots)$  from  $a(x)$ , and then divide the generating function by the expression  $x^k$ .
- Substitution of a polynomial  $f(x)$  for  $x$  leads to a specific combination of the terms of the original sequence. They can be expressed easily for  $f(x) = \alpha x$ , which corresponds to multiplication of the  $k$ -th term of the sequence by the scalar value  $\alpha^k$ . The substitution  $f(x) = x^n$  inserts  $n - 1$  zeros between each pair of adjacent terms.

The first and second rules express the fact that the assignment of the generating function to a sequence is an isomorphism of the two vector spaces (over the field in question).

There are other important operations which often appear when working with generating functions:

- *Differentiation* with respect to  $x$ : the function  $a'(x)$  generates the sequence  $(a_1, 2a_2, 3a_3, \dots)$ ; the term at index  $k$  is  $(k+1)a_{k+1}$  (i.e. the power series is differentiated term by term).
- *Integration*: the function  $\int_0^x a(t) dt$  generates the sequence  $(0, a_0, \frac{1}{2}a_1, \frac{1}{3}a_2, \frac{1}{4}a_3, \dots)$ ; for  $k \geq 1$ , the term at index  $k$  is equal to  $\frac{1}{k}a_{k-1}$  (clearly, differentiation of the corresponding power series term by term leads to the original function  $a(x)$ ).
- *Product of power series*: the product  $a(x)b(x)$  is the generating function for the sequence  $(c_0, c_1, c_2, \dots)$ , where

$$c_k = \sum_{i+j=k} a_i b_j,$$

i.e. the terms of the product are up to  $c_k$  the same as in the product  $(a_0 + a_1x + a_2x^2 + \dots + a_kx^k)(b_0 + b_1x + b_2x^2 + \dots + b_kx^k)$ . The sequence  $(c_n)$  is also called the *convolution of the sequences*  $(a_n), (b_n)$ .

**13.4.3. More links to continuous analysis.** There are useful examples of generating functions. Most of them are seen when working with power series in chapters five and six.

Perhaps the reader recognizes the generating function given by the geometric series:

$$a(x) = \frac{1}{1-x} = 1 + x + x^2 + \dots,$$

which corresponds to the constant sequence  $(1, 1, 1, \dots)$ . As we know, this power series converges for  $x \in (-1, 1)$  and equals  $1/(1-x)$ .

It works the other way round as well: Expand this function into its Taylor series at the point 0. The original series is obtained. This “encoding” of a sequence into a function and then decoding it back is the key idea in both the theory and practice of generating function.

Generally, consider any sequence  $a_i$  with  $\sqrt[n]{a_n}$  bounded. Then there is a neighbourhood on which its generating function converges (see 5.4.10 on page 428). For example, an easy check shows that this happens whenever  $|a_n| = O(n^k)$  with a constant exponent  $k \geq 0$ . On this neighbourhood, the generating functions can be worked with as with ordinary functions. In particular, one can add, multiply, compose, differentiate, and integrate them. All the equalities obtained carry over to the relevant sequences.

Recall several very useful basic power series and their sums:

$$\begin{aligned} \frac{1}{1-x} &= \sum_{n \geq 0} x^n, \\ \ln(1+x) &= \sum_{n \geq 1} (-1)^{n+1} \frac{x^n}{n}, \\ \ln \frac{1}{1-x} &= \sum_{n \geq 1} \frac{x^n}{n}, \\ e^x &= \sum_{n \geq 0} \frac{x^n}{n!}, \end{aligned}$$

$$\sin x = \sum_{n \geq 0} (-1)^n \frac{x^{2n+1}}{(2n+1)!},$$

$$\cos x = \sum_{n \geq 0} (-1)^n \frac{x^{2n}}{(2n)!}.$$

**13.4.4. Binomial theorem.** Recall the standard finite binomial formula  $(a+b)^r = a^r(1+c)^r = a^r \sum_{n=0}^r \binom{r}{n} c^n$ , where  $r \in \mathbb{N}$ ,  $0 \neq a, b \in \mathbb{C}$ ,  $c = b/a$ . Even if the power  $r$  is not a natural number, the Taylor series of  $(1+x)^r$  can still be computed. This yields the following generalization:



GENERALIZED BINOMIAL THEOREM

**Theorem.** For any  $r \in \mathbb{R}$ ,  $k \in \mathbb{N}$ , write

$$\binom{r}{k} = \frac{r(r-1)(r-2)\cdots(r-k+1)}{k!}$$

(in particular  $\binom{r}{0} = 1$ , having empty product divided by 1 in the latter formula). The power series expansion

$$(1+x)^r = \sum_{k \geq 0} \binom{r}{k} x^k$$

converges on a neighbourhood of zero, for each  $r \in \mathbb{R}$ . The latter formula is called the generalized binomial theorem.

In particular, the function  $\frac{1}{(1-x)^n}$ ,  $n \in \mathbb{N}$  can be expanded into the series

$$1 + \binom{1+n-1}{n-1} x + \cdots + \binom{k+n-1}{n-1} x^k + \cdots.$$

**PROOF.** The theorem is obvious if  $r \in \mathbb{N}$  since it is then the finite binomial formula. So assume  $r$  is not a natural number and thus zero is never obtained when evaluating  $\binom{r}{k}$ .

First, differentiate the function  $a(x) = (1+x)^r$  and evaluate all the derivatives in  $x = 0$ . Obviously

$$\begin{aligned} a^{(k)}(0) &= r(r-1)\cdots(r-k+1)(1+x)^{r-k} \Big|_{x=0} \\ &= r(r-1)\cdots(r-k+1) \end{aligned}$$

which provides the coefficients  $a_k = \binom{r}{k}$  of the series. In 5.4.5, there are several simple tests to decide about convergence of a number series. The ratio test helps here:

$$\frac{a_{k+1} x^{k+1}}{a_k x^k} = \frac{\frac{r(r-1)\cdots(r-k)}{(k+1)!}}{\frac{r(r-1)\cdots(r-k+1)}{k!}} x = \frac{r-k}{k+1} x.$$

By the ratio test, the radius of convergence is 1 for all  $r \notin \mathbb{N}$ .

The generalized binomial formula for negative integers is a straightforward consequence. If  $r = -n$ , we may write  $\binom{-n}{k} = (-1)^k \frac{1}{k!} (n+k-1)(n+k-2)\cdots n = \binom{n+k-1}{n-1}$  and substituting  $-x$  for the argument just kills the signs as requested.  $\square$

**13.4.5. Examples.** The formulae with  $r$  as a negative integer are very useful in practice. The simplest one is the geometric series with  $r = -1$ . Write down two more of them.



$$\frac{1}{(1-x)^2} = \sum_{n \geq 0} (n+1)x^n$$

$$\frac{1}{(1-x)^3} = \sum_{n \geq 0} \binom{n+2}{2} x^n.$$

The same results can be obtained by consecutive convolutions. Indeed, for the generating function  $a(x)$  of a sequence  $(a_0, a_1, a_2, \dots)$ ,  $\frac{1}{1-x}a(x)$  is the generating function for the sequence of all the partial sums  $(a_0, a_0+a_1, a_0+a_1+a_2, \dots)$ . For instance,

$$\frac{1}{1-x} \ln \frac{1}{1-x}$$

is the generating function of the *harmonic numbers*

$$H_n = 1 + \frac{1}{2} + \dots + \frac{1}{n}.$$

**13.4.6. Difference equations.** Typically, the generating functions can be very useful, if the sequences are defined by relations between their terms.



An instructive example of such an application is the complete discussion of the solutions of linear difference equations with constant coefficients. This is examined in the second part of chapter one, see 1.2.4. Back there, a formula is derived for first-order equations, the uniqueness and existence of the solution is justified, after only “guessing” the solution. Now, it can be truly derived in another straight forward way, working also in more complex and non-linear problems.

First, sort out the well-known example of the Fibonacci sequence, given by the recurrence

$$F_{n+2} = F_n + F_{n+1}, \quad F_0 = 0, \quad F_1 = 1,$$

and write  $F(x)$  for the (yet unknown) generating function of this sequence. We want to compute  $F(x)$  and so obtain an explicit expression for the  $n$ th Fibonacci number.

The defining equality can be expressed in terms of  $F(x)$  if we use our operations for shifting the terms of the sequence. Indeed,  $xF(x)$  corresponds to the sequence  $(0, F_0, F_1, F_2, \dots)$ , and  $x^2F(x)$  does to  $(0, 0, F_0, F_1, \dots)$ . Therefore, the generating function

$$G(x) = F(x) - xF(x) - x^2F(x)$$

represents the sequence

$$(F_0, F_1 - F_0, 0, 0, \dots, 0, \dots).$$

Substitute in the values  $F_0 = 0, F_1 = 1$  (the initial condition). Obviously  $G(x) = x$  and hence

$$(1 - x - x^2)F(x) = x.$$

$F(x)$  is a rational function. It can be rewritten as a linear combination of simple rational functions. This is helpful, since a

linear combination of generating functions corresponds to the same combination of the sequences.

Rational functions can be decomposed into partial fractions, see 6.2.7. Using this procedure, we find a generating function for  $1/(1 - x - x^2)$ . Namely, write  $r \in \mathbb{N}$

$$\begin{aligned} F(x) &= \frac{1}{1 - x - x^2} = \frac{A}{x - x_1} + \frac{B}{x - x_2} \\ &= \frac{a}{1 - \lambda_1 x} + \frac{b}{1 - \lambda_2 x}, \end{aligned}$$

where  $A, B$  are suitable (generally) complex constants, and  $x_1, x_2$  are the roots of the polynomial in the denominator. The ultimate constants  $a, b, \lambda_1$ , and  $\lambda_2$  can be obtained by a simple rearrangement of the particular fractions. This leads to the general solution for the generating function

$$F(x) = \sum_{n=0}^{\infty} (a\lambda_1^n + b\lambda_2^n)x^n,$$

and so the general solution of the recurrence is known as well.

In the present case, the roots of the quadratic polynomial are  $\frac{1 \pm \sqrt{5}}{2}$ . Hence the reciprocal values of the roots are  $\lambda_{1,2} = \frac{2}{1 \pm \sqrt{5}}$ . The partial fraction decomposition equality gives

$$x = a \left(1 - \frac{2}{1 - \sqrt{5}}x\right) + b \left(1 - \frac{2}{1 + \sqrt{5}}x\right)$$

and so  $a = -b = \frac{1}{\sqrt{5}}$ . Finally the requested solution

$$F_n = \frac{1}{\sqrt{5}} \left( \left(\frac{1 + \sqrt{5}}{2}\right)^n - \left(\frac{1 - \sqrt{5}}{2}\right)^n \right)$$

is obtained. Compare this procedure to the approach in 3.2.2 and 3.B.1. This expression, full of irrational numbers, is an integer. The second summand is approximately  $(1 - \sqrt{5})/2 \simeq -0.618$ . Its value is negligible for large  $n$ . Hence  $F_n$  can be computed by evaluating just the first summand and approximating to the nearest integer.

Of course, the same procedure can be applied for general  $k$ -th order homogeneous linear difference equations. Consider the recurrence

$$F_{n+k} = \alpha_0 F_n + \cdots + \alpha_{k-1} F_{n+k-1}.$$

The generating function for the resulting sequence is

$$F(x) = \frac{g(x)}{1 - \alpha_0 x^k - \cdots - \alpha_{k-1} x},$$

where the polynomial  $g(x)$  of order at most  $k - 1$  is determined by the chosen initial conditions.

Using partial fraction decomposition, the general result follows as in subsection 3.2.4.

**13.4.7. The general method.** Power series are a much stronger tool for solving recurrences. The point is that one is not restricted to linearity and homogeneity. Using the following general approach, recurrences that seem intractable at first sight can quite often be managed. The first steps are just algorithmic, while the final solution of the equation on the generating function may need very diverse approaches.





In order to be able to write down the necessary equations efficiently, adopt the convention of the *logical predicate*  $[\delta(n)]$  which is attached before the expression it should govern. Simply multiply by the coefficient 1 if  $\delta(n)$  is true, and by zero otherwise. For instance, the equation

$$F_n = F_{n-2} + F_{n-1} + [n = 1]1 + [n = 0]1$$

defines the above Fibonacci recurrence with initial conditions  $F_0 = 1$  and  $F_1 = 2$ .

METHOD TO RESOLVE RECURRENCES

Recurrent definitions of sequences  $(a_0, a_1, \dots)$  may be solved in the following 4 steps:

- (1) Write the complete dependence between the terms in the sequence as a single equation expressing  $a_n$  in terms of terms with smaller indices. This *universal formula* must hold for all  $n \in \mathbb{N}$  (supposing  $a_{-1} = a_{-2} = \dots = 0$ ).
- (2) Both sides of the equation are multiplied by  $x^n$ . Then sum the resulting expressions over all  $n \in \mathbb{N}$ . One of the summands is  $\sum_{n \geq 0} a_n x^n$ , which is the generating function  $A(x)$  for the sequence. Rearrange other summands so that they contain only the terms  $A(x)$  and some other polynomial expressions.
- (3) Solve the resulting equation with respect to  $A(x)$  explicitly.
- (4) The function  $A(x)$  is expanded into the power series. Its coefficients at  $x^n$  are the requested values of  $a_n$ .

As an example, consider a second order linear difference equation with constant coefficients, but a non-linear right hand side.

The recurrence is  $a_n = 5a_{n-1} - 6a_{n-2} - n$  with initial conditions  $a_0 = 0, a_1 = 1$ . The individual steps in the latter procedure are as follows:

*Step 1.* The universal equation is clear, up to the initial conditions. First check  $n = 0$ , which yields no extra term, but then  $n = 1$  enforces the extra value 2 to be added. Hence,

$$a_n = 5a_{n-1} - 6a_{n-2} - n + [n = 1]2.$$

*Step 2.*

$$\begin{aligned} \sum_{n \geq 0} a_n x^n &= 5x \sum_{n \geq 0} a_{n-1} x^{n-1} - 6x^2 \sum_{n \geq 0} a_{n-2} x^{n-2} \\ &\quad - \sum_{n \geq 0} n x^n + 2x \end{aligned}$$

Next, one of the terms is nearly the power series for  $(1-x)^{-2}$ . Thus remove one  $x$  there in order to get the equality on  $A(x)$  in the form as required (ignore the negative values of indices since all  $a_{-1}, a_{-2}, \dots$  vanish by assumption).

$$A(x) = 5xA(x) - 6x^2A(x) - x \frac{1}{(1-x)^2} + 2x.$$

*Step 3.* Find the reciprocal values to the roots, 2 and 3, of the polynomial  $1-5x+6x^2 = (1-2x)(1-3x)$ . An elementary

calculation yields

$$A(x) = \frac{2x^3 - 4x^2 + x}{(1-2x)(1-3x)(1-x)^2}.$$

*Step 4.* Partial fraction decomposition directly leads to the result

$$A(x) = -\frac{1}{4} \frac{1}{1-3x} + 2 \frac{1}{1-2x} - \frac{1}{2} \frac{1}{(1-x)^2} - \frac{5}{4} \frac{1}{1-x}.$$

This corresponds to the solution (notice the last but one term yields  $\sum_n (n+1)x^n$  and thus the constant term in our formula is the sum  $-1/2 - 5/4$ )

$$a_n = -\frac{1}{4} 3^n + 2^{n+1} - \frac{1}{2} n - \frac{7}{4}.$$

The first eight terms in the sequence are 0, 1, 3, 6, 8, -1, -59, -296, all integers of course.

### 13.4.8. Plane binary trees and Catalan numbers.



The next application of the generating functions answers the question about the number  $b_n$  of non-isomorphic plane binary trees on  $n$  vertices (cf. 13.1.18 for plane trees). Treat these trees in the form of the root (of a subtree) with a pair [*the left binary subtree, the right binary subtree*].

Examine the initial values of  $n$ , namely

$$b_0 = 1, b_1 = 1, b_2 = 2, b_3 = 5.$$

It is more or less obvious that for  $n \geq 1$ , the sequence  $b_n$  satisfies the recurrent formula

$$b_n = b_0 b_{n-1} + b_1 b_{n-2} + \cdots + b_{n-1} b_0,$$

and this is actually close to a convolution of two equal sequences. Rearrange the expression so that it holds for all  $n \in N_0$ :

$$b_n = \sum_{0 \leq k < n} b_k b_{n-k-1} + [n=0]1.$$

This finishes step 1 of the procedure.

In step 2, multiply both sides by  $x^n$  and add it all together. Write  $B(x)$  for the generating function of the sequence  $b_n$ .

$$\begin{aligned} B(x) &= \sum_{n,k} b_k b_{n-k-1} x^n + \sum_{n,k} [n=0] x^n \\ &= \sum_k b_k x^k \left( \sum_n b_{n-k-1} x^{n-k} \right) + 1 \\ &= \sum_k b_k x^k (xB(x)) + 1 = B(x) \cdot xB(x) + 1. \end{aligned}$$

Notice that the convolution  $b_n = b_0 b_{n-1} + b_1 b_{n-2} + \cdots + b_{n-1} b_0$  is replaced by

$$b_n = b_0 b_{n-1} + \cdots + b_{n-1} b_0 + b_n b_{-1} + b_{n+1} b_{-2} + \cdots.$$

This is no problem by the convention. It helps with processing the sums (it is much easier to work with infinite sums here than to keep an eye on the bounds all the time).

In step 3, the quadratic equation  $B(x) = xB(x)^2 + 1$  must be solved for  $B(x)$ . So

$$B(x) = \frac{1 \pm \sqrt{1 - 4x}}{2x}.$$

Although there are two solutions, not necessarily both must produce a valid solution to our problem. The sign  $+$  in the formula is impossible since then, the limit of  $B(x)$  for  $x \rightarrow 0_+$  is  $\infty$ , but the generating function for our sequence must have the value  $b_0 = 1$  at 0.

In the last step, expand  $B(x)$  into a power series. The expansion can be obtained using the generalized binomial theorem

$$\begin{aligned} (1 - 4x)^{1/2} &= \sum_{k \geq 0} \binom{1/2}{k} (-4x)^k \\ &= 1 + \sum_{k \geq 1} \frac{1}{2k} \binom{-1/2}{k-1} (-4x)^k. \end{aligned}$$

Dividing  $1 - \sqrt{1 - 4x}$  by the expression  $2x$  leads to

$$\begin{aligned} B(x) &= \sum_{k \geq 1} \frac{1}{k} \binom{-1/2}{k-1} (-4x)^{k-1} \\ &= \sum_{n \geq 0} \binom{-1/2}{n} \frac{(-4x)^n}{n+1} \end{aligned}$$

Substitute the  $(-4)^n$  multiple of  $\binom{-1/2}{n}$  into the definition of the generalized binomial numbers. A straightforward check shows  $(-4)^n \binom{-1/2}{n} = \binom{2n}{n}$ , which yields a final, much neater, formula for the coefficients. We conclude that the number of plane binary trees on  $n$  vertices equals

$$b_n = \frac{1}{n+1} \binom{2n}{n}.$$

These are known as the *Catalan numbers*. They occur surprisingly often:

- the number of well-parenthesized words of length  $2n$ , i.e. words consisting of  $n$  opening and  $n$  closing parentheses so that no prefix of the word contains more closing parentheses than closing ones;
- this also corresponds to the number of ways an unsupplied vending machine can accept  $n$  5-crown coins and  $n$  10-crown coins for 5-crown orders so that it can always give the change (hence the probability that a random ordering is satisfactory can also be found)
- the number of *monotonic* paths from  $[0, 0]$  to  $[n, n]$  along the sides of the unit squares of the grid such that the path does not cross the diagonal
- the number of triangulations of a convex  $(n + 2)$ -gon.

The intuitive reasoning for this is that they come from the expansion of the square root within  $B(x)$  and quadratic equalities appear often in real world.

**13.4.9. Quicksort analysis.** The next task is to determine the expected number of comparisons made by the Quicksort, a well-known algorithm for sorting a (finite) sequence of elements. This is the following divide and conquer type of algorithm:

PROCEDURE QSORT

*Input:* A (non-sorted) list of elements  $L = (L[0], \dots, L[n])$

*Output:* The sorted list  $L$  with the same elements.

- (1) if  $L$  is empty, then return the empty list  $()$ .
- (2) *Divide phase.* Create a sublist  $L_1$  by going through  $L$  and leaving only the elements  $x$  with  $L[0] > x$ , while putting the other elements into the list  $L_2$ .

- (3) *Conquer phase.* Combine the lists

$$L = \text{Qsort}(L_1) + (L[0]) + \text{Qsort}(L_2)$$

and return the list  $L$ .

We analyze how many comparisons are needed. Assume that all possible orderings of the list  $L$  to be sorted are distributed uniformly. The following parameters are crucial:

- The number of comparisons in the divide phase is  $n - 1$ .
- The assumption of uniformity ensures that the probability of  $L[0]$  being the  $k$ -th greatest element of the sequence is  $\frac{1}{n}$ .
- The sizes of the sublists to be sorted in the conquer phase are  $k - 1$  and  $n - k$ .

There is the following recurrent formula for the expected number of comparisons  $C_n$ :

$$(1) \quad C_n = n - 1 + \sum_{k=1}^n \frac{1}{n} (C_{k-1} + C_{n-k}).$$

One could work the steps of the general method directly, but the symmetry of the two summands allows a rewrite (1), multiplying by  $n$  at the same time

$$(2) \quad nC_n = n(n - 1) + 2 \sum_{k=1}^n C_{k-1}.$$

In the first step, care is needed concerning about  $n = 0$ . In the defining recurrence (1),  $n = 0$  is not treated at all (since the equation does not make sense). So the convention must be extended to include  $C_0 = 0$  in the computation. Then the equation (2) defines the  $C_1 = 0$  properly. It is not necessary to add any terms in view of the initial conditions.

Next, multiply both sides by  $x^n$  and add

$$\sum_{n \geq 0} nC_n x^n = \sum_{n \geq 0} n(n - 1)x^n + 2 \sum_{n \geq 0} \sum_{k=1}^n C_{k-1} x^n.$$

All the terms look familiar. The left hand side shows the derivative of the generating function  $C(x) = \sum_{n \geq 0} C_n x^n$  if one  $x$  is removed. The first term on the right is the series for  $(1 - x)^3$ , up to a constant and shift of powers by 2. Finally the last term is the convolution with  $(1 - x)^{-1}$ , up to one  $x$

and the coefficient 2. Hence the equation

$$xC'(x) = \frac{2x^2}{(1-x)^3} + 2\frac{x C(x)}{1-x}.$$

The third step is straightforward see 8.3.3. Divide by  $x$  to obtain

$$C'(x) = \frac{2}{1-x}C(x) + \frac{2x}{(1-x)^3}.$$

The corresponding integrating factor is  $e^{-\int \frac{2}{1-x} dx} = (1-x)^2$ . Hence

$$((1-x)^2 C(x))' = \frac{2x}{1-x},$$

and finally

$$C(x) = 2 \left( \frac{1}{(1-x)^2} \ln \frac{1}{1-x} - \frac{x}{(1-x)^2} \right).$$

The first terms in the bracket corresponds to the convolution of two known sequences, so it contributes to  $C_n$  by

$$\begin{aligned} \sum_{k=1}^n \frac{1}{k}(n-k+1) &= (n+1) \sum_{k=1}^n \frac{1}{k} - n \\ &= (n+1)H_n - n = (n+1)(H_{n+1} - 1), \end{aligned}$$

where  $H_n$  are the harmonic numbers. The result is

$$C_n = 2(n+1)(H_{n+1} - 1) - 2n.$$

Notice in 13.I.10, the very same recurrence is solved by different (more direct and simpler) tricks, without any differential equations involved.

Since the harmonic numbers  $H_n$  are easily approximated by  $\ln n = \int_1^n \frac{1}{x} dx$ , the analysis shows that the estimated time cost of quicksort is  $O(n \log n)$ . But it is easy to see that the worst time case is  $O(n^2)$  (in this version it happens if the list was already ordered properly – then  $L_1$  is always empty and the depth of the recursion is linear).

**13.4.10. Exponential generating functions.** Another approach



to generating functions is to take the exponential  $e^x = \sum_{n \geq 0} \frac{1}{n!} x^n$  as the power series corresponding to the constant sequence  $(1, 1, \dots)$ .

In general, this is called the *exponential generating functions*

$$\widehat{A}(x) = \sum_{n \geq 0} a_n \frac{x^n}{n!}.$$

Here are a few elementary examples:

$$\begin{aligned} e^x &\xleftrightarrow{\text{e.g.f.}} (1, 1, 1, \dots), \\ \frac{1}{1-x} &\xleftrightarrow{\text{e.g.f.}} (1, 1, 2, 6, 24, \dots, n!, \dots) \\ \ln \frac{1}{1-x} &\xleftrightarrow{\text{e.g.f.}} (0, 1, 1, 2, 6, 24, \dots) \end{aligned}$$

The slight modification of the definition (just reconsidering the  $\frac{1}{n!}$  coefficient) is responsible for a very different behaviour, compared to the ordinary generating functions. Indeed, now the elementary operations are:

- Multiplication of  $\widehat{A}(x)$  by  $x$  yields the sequence with terms  $\tilde{a}_n = na_{n-1}$ .
- Differentiation of  $\widehat{A}(x)$  shifts the sequence to the left.
- Integration of  $\widehat{A}(x)$  shifts the sequence to the right.
- The product of functions  $\widehat{A}(x)$  and  $\widehat{B}(x)$  corresponds to the sequence with terms  $h_n = \sum_k \binom{n}{k} a_k b_{n-k}$ , the *binomial convolution* of  $a_n$  and  $b_n$ .

As before, the exponential generating functions might become useful when resolving recurrences. Here is a simple example. Define the sequence by the initial conditions  $g_0 = 0$ ,  $g_1 = 1$  and the formula

$$g_n = -2ng_{n-1} + \sum_{k \geq 0} \binom{n}{k} g_k g_{n-k}.$$

At the first glance, seeing the binomial convolution suggests trying the exponential version.

Write  $\widehat{G}(x)$  for the corresponding power series and proceed in the usual four steps again.

*Step 1.* Complete the formula to accommodate the initial conditions:

$$g_n = -2ng_{n-1} + \sum_{k=0}^n \binom{n}{k} g_k g_{n-k} + [n = 1].$$

There seems to be a subtle point about  $g_0$  here, because the equation gives  $g_0 = g_0^2$ , with two solutions 0 and 1. The proper choice of  $g_0$  now yields the correct value for  $g_1$ , but the right solution  $\widehat{G}$  is chosen later.

*Step 2.* Multiply by  $\frac{x^n}{n!}$  and add over all  $n$ , to obtain

$$\widehat{G}(x) = -2x\widehat{G}(x) + \widehat{G}(x)^2 + x.$$

*Step 3.* Now, solve the easy quadratic equation, arriving at

$$\widehat{G}(x) = 1/2(1 + 2x \pm \sqrt{1 + 4x^2}).$$

The evaluation at zero provides  $g_0$ . Hence the right choice for  $g_0 = 0$  is the minus sign. Hence,

$$\widehat{G}(x) = \frac{1 + 2x - \sqrt{1 + 4x^2}}{2}.$$

*Step 4.* Apply the generalized binomial theorem, to expand  $\widehat{G}(x)$  into a power series. See 13.4.8.

$$\sqrt{1 + 4x^2} = 1 + \sum_{k \geq 1} \frac{1}{k} \cdot (-1)^{k-1} \cdot 2 \cdot \binom{2k-2}{k-1} \cdot x^{2k}.$$

Further, since

$$\widehat{G}(x) = \sum_{n \geq 0} g_n \frac{x^n}{n!} = \frac{1 + 2x - \sqrt{1 + 4x^2}}{2},$$

$g_{2k+1} = 0$  and

$$g_{2k} = (-1)^k \cdot \frac{1}{k} \binom{2k-2}{k-1} \cdot (2k)! = (-1)^k \cdot (2k)! \cdot C_{k-1},$$

where  $C_n$  is the  $n$ -th Catalan number.

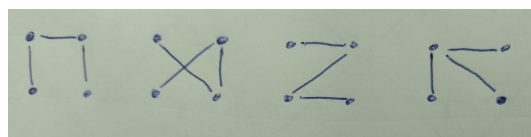
**13.4.11. Cayley's formula.** We conclude this chapter by a more complicated example.

Cayley's formula computes the number of trees (i.e. graphs with unique paths between all pairs of vertices) on  $n$  given vertices,

$$\kappa(K_n) = n^{n-2}.$$

The notation refers to the equivalent formulation to find all spanning trees in the complete graph  $K_n$ . Equivalently, in how many ways can a tree be realized on  $n$  vertices with the vertices labeled. For example, already the path  $P_n$  can be realized in  $n!$  ways, so there must be very many of them. This result is proved with the help of the exponential generating functions.

Write  $T_n = \kappa(K_n)$  for the unknown values. It is easily shown that  $T_1 = T_2 = 1$ ,  $T_3 = 3$ ,  $T_4 = 16$ . For instance, consider trees on 4 vertices. Out of the  $\binom{6}{3} = 20$  potential graphs with exactly three edges, those where the edges form a triangle must not be counted. There are  $\binom{4}{3} = 4$  of them. In the diagram, there are four different possibilities, and each of them can be rotated into another three, hence the solution is 16.



The recurrent formula can be obtained by fixing one of the vertices and add together the possibilities for all available degrees of this vertex. This suggests looking rather at the number  $R_n$  of the rooted trees. It is clear that  $R_n = nT_n$  because there are  $n$  possibilities to place the root at each of the trees. Also, one can work with one fixed ordering of the vertices in  $K_n$  and multiply the result by  $n!$  in the end. In this way, go through the possible degrees  $m$  of the first vertex and for each  $m$  to find the different possibilities for the sizes  $k_1 \dots, k_m$  of the corresponding subtrees. Obviously  $k_1 + \dots + k_m = n - 1$ , all  $k_i > 0$ , and since the labeling of all vertices is fixed, all the orders of the subtrees must be considered as equivalent. Multiply the contribution by  $\frac{1}{m!}$  and similarly for each of the possibilities of the subtrees. The recurrent formula is

$$R_n = n! \sum_{m>0} \frac{1}{m!} \sum_{k_1+\dots+k_m=n-1} \frac{1}{k_1! \dots k_m!} R_{k_1} \dots R_{k_m}.$$

Of course,  $R_0 = 0$ ,  $R_1 = 1$  and, already using the formula,  $R_2 = 2u_1 = 2$ . Next,  $R_3 = 3u_2 + 3u_1^2 = 9$ ,  $R_4 = 4R_3 + 24R_1R_2 + 4R_1^3 = 64$ , all as expected. The first step of the standard procedure is accomplished.

Next, write  $\widehat{R}(x) = \sum_{n \geq 0} R_n \frac{1}{n!} x^n$ . The inner sum is the coefficient at  $x^{n-1}$  in the  $m$ -th power of the series  $\widehat{R}(x)$ . Therefore,

$$R_n \frac{1}{n!} = [x^{n-1}] \sum_{m \geq 0} \frac{1}{m!} \widehat{R}(x)^m,$$

and hence have the required equation on  $\widehat{R}$ :

$$\widehat{R}(x) = x e^{\widehat{R}(x)}.$$

There are several ways, of solving such functional equations. Here is one such tool without proof.

maybe, get it as application of residue theorem in chapter 9

**Theorem** (Lagrange inverse formula). *Consider an analytic function  $f$ ,  $f(0) = 0$  and  $f'(0) \neq 0$ . Then there (locally) is the analytic inverse of  $f$ , i.e  $w = g(z) = \sum_{n \geq 1} g_n \frac{z^n}{n!}$  and  $z = f(g(z))$ . Moreover, for all  $n > 0$ ,*

$$g_n = \lim_{w \rightarrow 0} \left( \frac{d^{n-1}}{dw^{n-1}} \left( \frac{w}{f(w)} \right)^n \right).$$

In this case, solve the equation  $x = \frac{\widehat{R}(x)}{e^{\widehat{R}(x)}}$ , so that we may apply the latter theorem with  $g = \widehat{R}$  and  $f(w) = \frac{w}{e^w}$ . It follows that

$$\begin{aligned} [x^n] \widehat{R}(x) &= \frac{1}{n} [w^{n-1}] \left( \frac{w}{w/e^w} \right)^n \\ &= \frac{1}{n} [w^{n-1}] e^{wn} = \frac{1}{n} \frac{n^{n-1}}{(n-1)!} = \frac{n^{n-1}}{n!} \end{aligned}$$

In particular,  $R_n = n^{n-1}$  and so,

$$T_n = \frac{R_n}{n} = n^{n-2}.$$



**J. Additional exercises to the whole chapter**

**13.J.1.** Determine the number of edges that must be added into

- i) the cycle graph  $C_n$  on  $n$  vertices,
- ii) the complete bipartite graph  $K_{m,n}$

in order to obtain a complete graph. ○

**13.J.2.** Let the vertices of  $K_6$  be labeled  $1, 2, \dots, 6$  and let every edge  $\{i, j\}$  be assigned the integer  $[(i + j) \bmod 3] + 1$ . How many maximum spanning trees are there in this graph? ○

**13.J.3.** Let the vertices of  $K_7$  be labeled  $1, 2, \dots, 7$  and let every edge  $\{i, j\}$  be assigned the integer  $[(i + j) \bmod 3] + 1$ . How many maximum spanning trees are there in this graph? ○

**13.J.4.** Let the vertices of  $K_5$  be labeled  $1, 2, \dots, 5$  and let every edge  $\{i, j\}$  be assigned the integer: 1 if  $i + j$  is odd; 2 if  $i + j$  is even. How many maximum spanning trees are there in this graph? ○

**13.J.5.** Let the vertices of  $K_5$  be labeled  $1, 2, \dots, 5$  and let every edge  $\{i, j\}$  be assigned the integer: 1 if  $i + j$  is odd; 2 if  $i + j$  is even. How many minimum spanning trees are there in this graph? ○

**13.J.6.** Let the vertices of  $K_6$  be labeled  $1, 2, \dots, 6$  and let every edge  $\{i, j\}$  be assigned the integer: 1 if  $i + j$  leaves remainder 1 upon division by 3; 2 if  $i + j$  leaves remainder 2 upon division by 3; 3 if  $i + j$  is divisible by 3; How many minimum spanning trees are there in this graph? ○

**13.J.7.** Let the vertices of  $K_6$  be labeled  $1, 2, \dots, 6$  and let every edge  $\{i, j\}$  be assigned the integer: 1 if  $i + j$  leaves remainder 1 upon division by 3; 2 if  $i + j$  leaves remainder 2 upon division by 3; 3 if  $i + j$  is divisible by 3; How many maximum spanning trees are there in this graph? ○

**13.J.8. Icosian Game** – find a Hamiltonian cycle in the graph consisting of the vertices and edges of the regular dodecahedron. □

**Solution.** See Wikipedia<sup>4</sup>. □

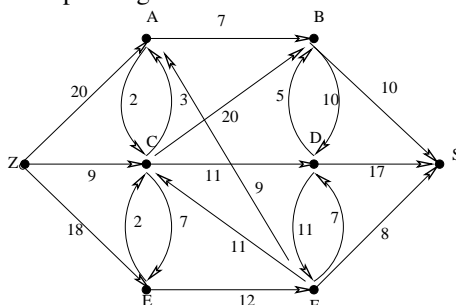
**13.J.9.** Does there exist a Hamiltonian cycle in the Petersen graph?

**Solution.** No (however, when any one of the vertices is removed, the resulting graph is already Hamiltonian). This can be shown by enumerating all 3-regular Hamiltonian graphs on 10 vertices and finding a cycle of length less than 5 in each of them. □

**13.J.10.** If  $G = (V, E)$  is Hamiltonian and  $\emptyset \neq W \subsetneq V$ , then  $G \setminus W$  has at most  $|W|$  connected components. Give an example of a graph where the converse does not hold.

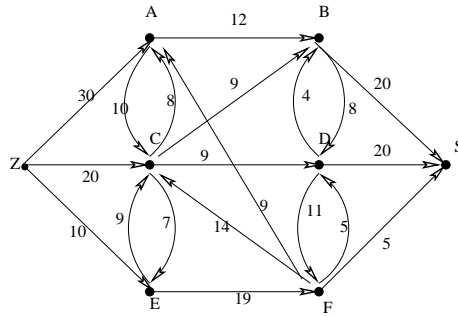
**Solution.** □

**13.J.11.** Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:

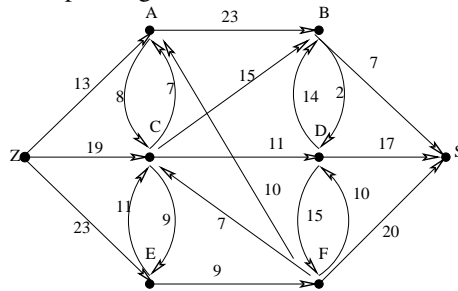


**13.J.12.** Find a maximum flow and the corresponding minimum cut in the following weighted directed graph: ○

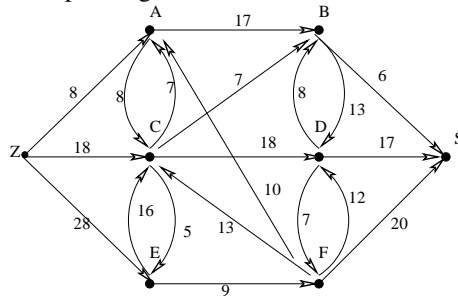
<sup>4</sup>Wikipedia, *Icosian game*, [http://en.wikipedia.org/wiki/Icosian\\_game](http://en.wikipedia.org/wiki/Icosian_game) (as of Aug. 8, 2013, 13:24 GMT).



13.J.13. Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:



13.J.14. Find a maximum flow and the corresponding minimum cut in the following weighted directed graph:



13.J.15. Find the generating functions of the following sequences:

- i)  $(1; 2; 1; 4; 1; 8; 1; 16; \dots)$
- ii)  $(1; 1; 0; 1; 1; 0; 1; 1; \dots)$
- iii)  $(1; -1; 2; -2; 3; -3; 4; -4; \dots)$

**Solution.**

i)  $(1; 2; 1; 4; 1; 8; 1; 16; \dots) = (1; 0; 1; 0; \dots) + (0; 2; 0; 4; 0; 16; \dots)$ . Thus, we find the generating functions for each sequence separately. As for the first one, consider the sequence  $(1, 1, 1, 1, \dots)$ . It is generated by the function  $\frac{1}{1-x}$ .

The zeros can be inserted by substituting  $x^2$  for  $x$ . As for the second sequence, we proceed similarly, starting with  $(1; 2; 4; 8; 16; \dots)$ , then multiplying by two, inserting zeros, and finally shifting to the right by multiplying by  $x$ .

ii)  $(1; 1; 0; 1; 1; 0; 1; 1; \dots) = (1; 0; 0; 1; 0; 0; 1; \dots) + (0; 1; 0; 0; 1; 0; 0; 1; \dots)$ .

i)  $\frac{1}{1-x^2} + \frac{2x}{1-2x^2}$

ii)  $\frac{1+x}{1-x^3}$

iii)  $\frac{-1}{(1-x^2)^2} + \frac{x}{(1-x^2)^2}$

□

**13.J.16.** Find the coefficient at  $x^{17}$  in  $(x^3 + x^4 + x^5 + \dots)^3$ .

**Solution.**  $(x^3 + x^4 + x^5 + \dots)^3 = \frac{x^9}{(1-x)^3} = x^9 \cdot \frac{1}{(1-x)^3}$ . We are thus looking for the coefficient at  $x^8$  in  $\frac{1}{(1-x)^3}$ . This is equal to  $\binom{10}{2}$ , i. e. 45.  $\square$

**13.J.17.** There are 30 red, 40 blue, and 50 white balls in a box (balls of the same color are indistinguishable). In how many ways can we pick up 70 balls from the box?  $\circ$

**Solution.** Clearly, the number of possibilities is equal to the coefficient at  $x^{70}$  in the expression

$$(1 + x + \dots + x^{30})(1 + x + \dots + x^{40})(1 + x + \dots + x^{50}).$$

Mere rearrangements lead to

$$(1+x+\dots+x^{30})(1+x+\dots+x^{40})(1+x+\dots+x^{50}) = \frac{1}{(1-x)^3} \dots (1-x^{31})(1-x^{41})(1-x^{51}).$$

Applying the generalized binomial theorem, we obtain the solution  $\binom{72}{2} - \binom{41}{2} - \binom{31}{2} - \binom{21}{2}$ .  $\square$

**13.J.18.** What is the probability that a roll of 12 dice results in the sum of 30?

*Hint: Express the number of possibilities when the sum is 30. Consider  $(x + x^2 + x^3 + x^4 + x^5 + x^6)^{12}$ .*  $\circ$

**13.J.19.** A fruit grower wants to plant 25 new trees, having four species at his disposal. However, his wife insists that there be at most 1 walnut, at most 10 apples, at least 6 cherries, and at least 8 plums. In how many ways can he fulfill his beloved's wishes?

*Hint: We are interested in the coefficient at  $x^{25}$  in the expression*

$$(1+x)(1+x+\dots+x^{10})(x^6+x^7+\dots)(x^8+x^9+\dots).$$

**Solution.**

$$(1+x)(1+x+\dots+x^{10})(x^6+x^7+\dots)(x^8+x^9+\dots) = \frac{x^{14}(1-x^2)(1-x^{11})}{(1-x)^4}.$$

Therefore, we are looking for the coefficient at  $x^{11}$  in  $(1-x^2-x^{11}\dots) \cdot \frac{1}{(1-x)^4}$ , which is equal to  $\binom{14}{3} - \binom{12}{3} - \binom{3}{3}$ .  $\square$

**13.J.20.** Express the general term of the sequences defined by the following recurrences:

i)  $a_1 = 3, a_2 = 5, a_{n+2} = 4a_{n+1} - 3a_n$  for  $n = 1, 2, 3, \dots$

ii)  $a_0 = 0, a_1 = 1, a_{n+2} = 2a_{n+1} - 4a_n$  for  $n = 0, 1, 2, 3, \dots$

**Solution.**

i)  $a_n = 2 + 3^{n-1}$ .

ii)  $a_n = \frac{1}{2}\sqrt{-3} \cdot ((1 + \sqrt{-3})^n - (1 - \sqrt{-3})^n)$ .

$\square$

**13.J.21.** Solve the recurrence where each term of the sequence  $(a_0, a_1, a_2, \dots)$  is equal to the arithmetic mean of the preceding two terms.  $\circ$

**13.J.22.** Solve the recurrence  $a_{n+2} = \sqrt{a_{n+1}a_n}$  with the initial conditions  $a_0 = 2, a_1 = 8$ .

*Hint: Create a new sequence  $b_n = \log_2 a_n$ .*  $\circ$

**13.J.23.** Solve the recurrence given by

$$a_n = \sum_{k \geq 0} \binom{n}{k} \frac{a_k}{2^k}, a_0 = 1.$$

*Hint: Multiply both sides by  $\frac{x^n}{n!}$  and sum it up. Note that  $\widehat{A}(X)$  is the exponential generating function for the sequence  $(a_n)$ .*  $\circ$

**13.J.24.** Find the number of triangulations of a convex  $n$ -gon.

*Hint: Select any diagonal that goes through a fixed vertex, this splits the polygon in two.*

**Solution.**  $t_n = C_{n-2}$ , where  $C_n$  denotes the  $n$ -th Catalan number. □

**13.J.25.** Find the number of walks in a square grid of size  $n \times n$  from the lower left-hand corner  $A$  to the upper right-hand corner  $B$  which go only upwards or rightwards and intersect the diagonal  $AB$  at exactly one point (besides  $A$  and  $B$ ).

*Hint: Catalan numbers.* ○

**13.J.26.** Prove that the Fibonacci number satisfy:

i)  $F_2 + F_4 + \cdots + F_{2n} = F_{2n+1} - 1$

ii)  $F_1 + F_3 + \cdots + F_{2n-1} = F_{2n}$

○

**13.J.27.** Recall the well-known puzzle Tower of Hanoi and let  $H_n$  denote the minimum number of steps necessary to move a tower consisting of  $n$  disks from one rod to another one. Find a recurrent formula for  $H_n$  as well as its general solution.

**Solution.**  $H_{n+1} = 2H_n + 1$ ,  $H_n = 2^n - 1$ . □

At the very end of the book, we present one problem from practice.

- 13.J.28.** A volleyball team (with a libero, i. e. 7 people) are sitting in a pub, drinking their favorite and well-deserved beer. However, there are only 7 beer mugs available. What is the probability that in the next round, i) exactly one volleyball player is not given the same mug as the last time,  
 ii) no one is given the same mug as the last time,  
 iii) exactly three players are given the same mug.

**Solution.**

- i) If six of the seven people are given the same mug, then so must the last one. Therefore, the probability is zero.  
 ii) Let  $M$  be the set of all orders of the 7 players and let  $A_i$  be the event of orders where the  $i$ -th player is given his mug. We want to calculate  $|M - \cup_i A_i|$ . We get  $7! \sum_{k=0}^7 \frac{(-1)^k}{k!} = 1854$ , so the probability is  $\frac{1854}{5040} = \frac{103}{280} \doteq 0,37$ .  
 iii) There are  $\binom{7}{3} = 35$  ways to select the three people who are to get the same mug. The remaining four people must be given different mugs. Again, we can apply the formula from above, i. e., there are  $4! \sum_{k=0}^4 \frac{(-1)^k}{k!} = 9$  possibilities. Altogether, there are  $9 \cdot 35 = 315$  favorable cases, so the probability is  $\frac{315}{5040} = \frac{1}{16}$ .



□

Key to the exercises

**13.A.23.** The cut vertices are 0, 1, 9, 10; the cut edges are (0, 1), (0, 12), (9, 10).

**13.B.3.** (3, 1), (3, 2), (3, 4), (3, 5), (3, 6), (6, 1), (6, 2), (6, 4), (6, 5), (5, 1), (5, 2), (5, 4), (4, 1), (4, 2), (2, 1).

**13.B.4.** (3, 1), (3, 2), (3, 4), (3, 5), (3, 6), (1, 2), (1, 4), (1, 5), (1, 6), (2, 4), (2, 5), ... (5, 6).

**13.B.5.** Solution is missing.

**13.B.12.**

It can be shown using the Havel–Hakimi procedure that such graph indeed exists. However, it cannot be planar:  $|V| = 10$ ,  $|E| = 35$ , but if it were planar, we would have  $3|V| - 6 \geq |E|$ , i. e.,  $24 \geq 35$ .

**13.B.15.**

- i) Yes. This follows immediately from the Kuratowski theorem ( $K_5$  has 10 edges and  $K_{3,3}$  has 9).
- ii) No. Consider  $K_5$  or  $K_{3,3}$ .
- iii) No. There are many counterexamples, for instance  $K_{3,3}$  with another vertex and an edge leading to it.
- iv) No. Consider  $K_5$ .
- v) No. Consider  $K_{3,3}$ .
- vi) The same as (ii).
- vii) No. Consider  $C_n$ .
- viii) No. Consider  $K_5$ .
- ix) No. Consider  $C_n$ .

**13.B.17.** The first code does not represent a tree (it has a proper prefix with the same number of zeros and ones). There is a tree corresponding to the second code.

**13.C.4.** The procedure is incorrect. As a counterexample, consider a cycle graph with one edge of length two and all other edges of length one.

**13.C.5.** Applying any algorithm for finding a minimum spanning tree, we find out that the wanted length is 12154 (the spanning tree consists of edges LPe, LP, LNY, PeT, MCNY).

**13.C.6.** Solution is missing.

**13.D.5.** We find a maximum flow of size 15 and the cut [1, 6], [1, 3], [2, 4], [2, 3] of the same capacity.

**13.D.7.** We know from the theory and the result of the above exercise that the minimum capacity of a cut is 9. There are more maximum flows in the network. For instance, we can set  $f(a) = 2$ ,  $f(b) = 4$ ,  $f(c) = 1$ ,  $f(h) = 1$ ,  $f(j) = 4$ ,  $f(f) = 2$ ,  $f(i) = 7$ , and  $f(v) = 0$  for all other edges  $v$  of the graph.

**13.E.5.** Solution is missing.

**13.E.11.**

$$1 - \frac{4! \cdot 4!}{2^4} = \frac{27}{35}$$

**13.E.12.**  $\frac{49}{54}$ .

**13.I.4.**

- i) We know from the exercise of subsection 13.4.3 that the generating function of the sequence (1, 2, 3, 4, ...) is  $\frac{1}{(1-x)^2}$ .
- ii) Since we have (by the previous exercise as well)

$$\frac{x}{(1-x)^2} \xleftrightarrow{\text{o.g.f.}} (0, 1, 2, 3, \dots),$$

mÄĀme pro derivaci tÄĀsto funkce

$$\left( \frac{x}{(1-x)^2} \right)' = \frac{1+x}{(1-x)^3} \xleftrightarrow{\text{o.g.f.}} (1 \cdot 1, 2 \cdot 2, 3 \cdot 3, \dots).$$

Let us emphasize that this problem could also be solved using the fact that  $\frac{1}{(1-x)^3} \xleftrightarrow{\text{o.g.f.}} \binom{n+2}{n}$ .

iii) We have

$$\begin{aligned} \frac{1}{1-x} &\overset{\text{o.g.f.}}{\longleftrightarrow} (1, 1, 1, 1, \dots), \\ \frac{1}{1-2x} &\overset{\text{o.g.f.}}{\longleftrightarrow} (1, 2, 4, 8, \dots), \\ \frac{1}{1-2x^2} &\overset{\text{o.g.f.}}{\longleftrightarrow} (1, 0, 2, 0, 4, 0, \dots), \\ \frac{x}{1-2x^2} &\overset{\text{o.g.f.}}{\longleftrightarrow} (0, 1, 0, 2, 0, 4, \dots), \end{aligned}$$

whence we get the result

$$\frac{1+x}{1-2x^2} \overset{\text{o.g.f.}}{\longleftrightarrow} (1, 1, 2, 2, 4, 4, 8, 8, \dots).$$

iv) We know from the above that  $f(x) = \frac{1+x}{(1-x)^3} \overset{\text{o.g.f.}}{\longleftrightarrow} (1^2, 2^2, 3^2, \dots)$ , hence

$$\frac{f(x) - (1+4x)}{x^2} \overset{\text{o.g.f.}}{\longleftrightarrow} (3^2, 4^2, 5^2, \dots).$$

Substituting  $2x^3$  for  $x$ , we obtain

$$\frac{f(2x^3) - (1+8x^3)}{4^6} \overset{\text{o.g.f.}}{\longleftrightarrow} (9, 0, 0, 2 \cdot 16, 0, 0, 4 \cdot 25, \dots).$$

v) If we denote the result of the previous problem as  $F(x)$ , then the result of this one is

$$F(x) - x^2 F(x) + \frac{x}{1-x^3}.$$

**13.I.11.**  $x/(1-3x+x^2)$ .

**13.J.1.**

- i) The complete graph on  $n$  vertices has  $\frac{n(n-1)}{2}$  edges, the cycle graph on  $n$  vertices has  $n$  edges. Therefore,  $\frac{n(n-1)}{2} - n$  edges must be added to the cycle graph.
- ii) Similarly as above, we get the result  $\frac{(m+n)(m+n-1)}{2} - m \cdot n$ .

**13.J.2.** There are five edges whose value is 3: four of them lie on the cycle 23562 and the remaining one is the edge 14. Therefore, they form a disconnected subgraph of the complete graph, so the spanning tree must contain at least one edge of value 2. Thus, the total weight of a maximum spanning tree is at most  $4 \cdot 3 + 2 = 14$ . And indeed, there exist spanning trees with this weight. We select all the edges of value 3 except for one that lies on the mentioned cycle and connect the resulting components 2356 and 14 with any edge of value 2. There are four such edges. Altogether, there are  $4 \cdot 4 = 16$  maximum spanning trees.

**13.J.3.** The edges of value 1 form a subgraph with two connected components, namely  $\{1, 2, 4, 5, 7\}$  a  $\{3, 6\}$ . Further, there are six edges of value 2 that lead between these two components. Therefore, the total weight of a minimum spanning tree is  $6 \cdot 1 + 2 = 8$ . Moreover, there are exactly three cycles in the former component, each of length 4, and each of the 6 edges of this component belongs to exactly two of the three cycles. In order to obtain a tree from this component, we must omit two edges, which can be done in  $6 \cdot 4/2$  ways. Altogether, we get  $12 \cdot 6 = 72$  minimum spanning trees.

**13.J.4.** 18.

**13.J.5.** 12.

**13.J.6.** 16.

**13.J.7.** 16.

**13.J.11.** The minimum cut is given by the set  $\{Z, A, E\}$ . Its value is 32.

**13.J.12.** The minimum cut is given by the set  $\{B, D, S\}$ . Its value is 40.

**13.J.13.** The minimum cut is given by the set  $\{F, S, D\}$ . Its value is 29.

**13.J.14.** The minimum cut is given by the set  $\{F, S\}$ . Its value is 39.

**13.J.18.** The resulting probability is the ratio of the number of favorable cases to the number of all cases. Clearly, the latter is  $6^{12}$ . Now, let us compute the number of favorable cases. Consider the expression  $(x + x^2 + x^3 + x^4 + x^5 + x^6)^{12}$ . Then, the number of favorable cases is the coefficient at  $x^{30}$ . We have:

$$(x+x^2+x^3+x^4+x^5+x^6)^{12} = \left(\frac{x(1-x^6)}{1-x}\right)^{12} = x^{12} \cdot \left(\frac{1-x^6}{1-x}\right)^{12}$$

Therefore, we are interested in the coefficient at  $x^{18}$  in

$$\left(\frac{1-x^6}{1-x}\right)^{12} = (1-12x^6+66x^{12}-220x^{18}) \cdot \frac{1}{1-x}^{12}.$$

It follows from the generalized binomial theorem that the number of favorable cases is

$$\binom{29}{11} - 12 \cdot \binom{23}{11} + 66 \cdot \binom{17}{11} - 220 \cdot \binom{11}{11}.$$

**13.J.21.**  $a_n = k \left(-\frac{1}{2}\right)^n + l$ .

**13.J.22.** Solution is missing.

**13.J.23.** Solution is missing.

**13.J.25.** Solution is missing.

**13.J.26.** Solution is missing.



---

## Index

- LU*-decomposition, 172
  - $\delta$ -neighborhood, 264
  - $\varepsilon$ -net, 457
  - $k$ -times (continuously) differentiable, 481
  - $k$ -combination, 13
  - $k$ -edge-connected, 634
  - $k$ -linear, 107
  - $k$ -permutation, 13
  - $k$ -permutations, 15
  - $k$ -vertex-connected, 633
  - (Riemann) measurable, 373
  - Taylor expansion with a remainder*, 345
  - characteristic polynomial of the mapping*, 111
  - commutative group*, 5
  - commutative rings*, 5
  - degree of nilpotency*, 163
  - factor vector space*, 166
    - generated*, 204
    - isomorphism*, 99
  - linear combination*, 93
  - linearly dependent*, 93
  - linearly independent*, 93
  - matrix for changing the basis*, 102
  - matrix of the mapping  $f$* , 101
  - normal*, 161
  - normal matrices*, 162
  - parallelepiped*, 209
  - root subspace*, 165
  - scalars*, 6
  - Sylvester criterion, 228
- angle between vector subspaces, 215
- A binary relation, 37
- a critical point of order  $k$ , 349
- a Laurent series centered at  $x_0$ , 383
- a representative, 40
- Abel's theorem, 383
- absolute value, 260
- absolutely convergent, 292
- absorbing, 152
- addition inverse, 74
- adjacency list, 629
- adjacency matrix, 629
- adjacent, 622
- adjoint mapping, 159
- adjoint matrix, 159
- admissible vector  $x$ , 136
- affine combinations of points, 207
- affine coordinate system, 204
- affine coordinates, 26
- affine frame, 204
- affine hull, 205
- affine map, 211
- affine mappings of the plane, 29
- affine subspace, 204
- algebra, 249
- algebraic complement, 89
- Algebraic multiplicity, 115, 163
- algebraically adjoint matrix, 91
- algorithm ElGamal, 614
- alternating series, 294
- an asymptote with a slope, 351
- an inflective point, 350
- analytic formula, 223
- analytic function, 347
- angle  $\varphi(u, v)$ , 215
- angle between affine subspaces, 216
- angles, 208
- antisymmetric, 39, 86
- antisymmetric bilinear form, 108
- area, 35
- arithmetic mean, 288
- arithmetic representatives, 230
- associativity, 5
- asymptotes, 351
- asymptotes without a slope, 351
- asymptotic estimation, 358
- autonomous systems of ordinary differential equations, 538
- axioms, 5
- axis of collineation  $f$ , 234
- Baire space, 455
- balanced, 627, 642
- basis, 24
- basis of vector space, 96
- Bernoulli differential equations, 529
- Bernoulli's inequality, 289
- Bessel inequality, 156
- Bessel's identity, 421
- Bessel's inequality, 420
- Bezout's identity, 568
- bilinear forms, 107
- binary search trees, 642
- binary trees, 642
- binomial coefficient, 13
- binomial congruences, 594
- binomial expansion, 13
- binormal of the curve, 356
- bipartite matching, 657
- Borůzavka's algorithm, 652
- boundary, 208
- boundary point, 264, 456
- bounded, 136, 264, 455, 457

- breadth-first search, 632
- bunch of hyperplanes passing through point  $A \in \mathcal{P}(V)$ , 234
  
- calibre, 388
- canonical analytic formulas, 224
- capacities, 654
- capacity of the cut, 655
- Cartesian coordinate system, 213
- Cartesian product, 37
- Casortian, 139
- Cauchy inequality, 155
- Cauchy sequence, 261, 446
- Cauchy theorem, 88
- Cauchy's mean value theorem, 285
- center, 225
- center of collineation  $f$ , 234
- characteristic function of the set, 373
- characteristic polynomial, 140
- characteristic polynomial of the matrix, 110
- child, 642
- classical probability, 17
- classical Stokes' theorem, 524
- clique, 626
- closed, 446
- closed interval, 263
- closed subset, 263
- closure, 263
- code of a plane tree, 643
- codomain, 7
- codomain of the relation, 37
- coefficients of the polynomial, 250
- collineation, 231
- coloring of the graph, 625
- columns of the matrix, 74
- combinations, 13
- common divisor, 567
- commutative field, 5
- commutativity, 5
- compact, 264, 456
- comparable, 39
- complementary event, 18
- complementary minor, 89
- complementary submatrix, 89
- complete, 39, 262, 541
- complete bipartite graph, 623
- complete graph, 623
- complete metric spaces, 446
- complete orthogonal system, 421
- complete residue system, 581
- complete spline, 257
- completion of the metric space  $X$ , 451
- complex Fourier coefficients, 424
- complex-valued functions of a real variable, 249
- composite, 570
- composition, 38
- composition of a relation, 38
- concave, 350
- conditional probability, 21
- congruence in variable, 588
- congruent modulo  $m$ , 577
- conjugation, 260
- connected, 633
- connected components, 633
- consecutive, 622
- continuous at a point, 272
- continuous mappings, 447
  
- contraction, 531
- contraction mapping, 454
- convergence, 261
- converges, 261, 446
- converges uniformly, 379
- convex, 208, 350
- convex hull, 208
- convex polyhedrons, 209
- convolution, 664
- convolution of functions, 436
- coordinates of the vector, 99
- coprime, 570
- cosine Fourier series, 427
- critical points, 344
- cross product, 221
- cross-ratio of four points  $(A, B, C, D)$ , 233
- crossbar, 210
- cubic interpolation spline, 257
- curvature, 353
- cut in a network, 655
- cycle, 86, 625
- cycle graph of length  $n$ , 623
- cycle ladder graph, 624
- cycle of length  $n$ , 625
- cyclic, 163
- Cyclometric functions, 299
  
- $d$ , 627
- Darboux integral, 366
- decomposition into partial fractions, 363
- deconvolution, 441
- decreasing at a point, 279
- decreasing on an interval, 279
- definite integral, 359
- degree  $n$ , 250
- degree sequence, 627
- deleted neighborhood, 267
- dense, 451
- dependent variable, 7
- depth-first search, 632
- derivative, 254, 277
- derivative in the direction of a vector, 476
- determinant, 29
- determinant expansion, 88
- Determinant of the matrix  $A$ , 84
- diameter of the set, 455
- difference equation of the first order, 9
- difference space, 203
- differentiable, 277
- differentiable  $(k + 1)$  times, 342
- differentiable at a point, 477
- differentiable mapping, 488
- differential equations, 370
- differential of the function  $f$ , 477
- differential of the mapping, 489
- Dijkstra's algorithm, 636
- dimension, 72, 203
- dimension of  $V$ , 96
- dimension of the matrix, 75
- Dirac function  $\delta$ , 441
- direct predecessor, 642
- direct successor, 642
- direct sum, 95
- directed graph, 622
- directional derivative, 476
- Dirichlet condition, 431

- Dirichlet kernel, 433  
 Dirichlet product, 579  
 discontinuities, 369  
 discrete logarithm, 584  
 discrete logarithm problem, 614  
 discrete logistic model, 12  
 distance, 457  
 distance of the point  $x$  from the set  $A$ , 263  
 distance of vertices, 635  
 distributivity of addition with respect to multiplication, 5  
 diverges, 291  
 divisible, 566  
 domain, 7  
 double differentiable, 342  
 dual basis, 103  
 dual mapping, 158  
 dual projective space, 233  
 dual space, 103  
  
 eccentricity, 644  
 echelon form, 78  
 edge list, 629  
 edges, 622  
 eigenvalues of mapping, 110  
 eigenvalues of the matrix, 110  
 eigenvectors of mapping, 110  
 elementary column transformations, 78  
 elementary row transformations, 78  
 endpoints, 622  
 equicontinuous, 460  
 equivalence, 577  
 equivalence classes, 40  
 equivalence relation, 39  
 Euclid's algorithm, 568  
 Euclidean distance, 213  
 Euclidean division, 566  
 euclidean point space, 213  
 Euclidean subspaces, 213  
 euclidean transformation, 225  
 Euler's approximation, 533  
 Euler's number  $e$ , 290  
 Eulerian, 638  
 Eulerian trail, 638  
 even, 85  
 event, 17  
 events, 17  
 exponential function, 275  
 exponential generating functions, 669  
 exponential of a matrix, 545  
 exterior differential, 521  
 exterior differential  $k$ -form, 515  
  
 faces, 208  
 faces of the planar graph, 646  
 factor, 626  
 father wavelet function, 428  
 Fermat numbers, 609  
 field, 5  
 finite automaton, 623  
 finitely dimensional, 96  
 first order recurrence, 9  
 fixed hyperplane, 234  
 fixed point, 234  
 flow, 654  
 flow of the vector field, 541  
 Ford-Fulkerson algorithm, 655  
 forest, 640  
  
 Fourier coefficients, 421  
 Fourier coefficients of the function, 422  
 Fourier cosine transform, 443  
 Fourier series, 421, 422  
 Fourier sine transform, 443  
 Fourier transform, 438  
 Fourier transform  $\mathcal{F}$ , 439  
 free vectors, 25  
 Frenet–Serret formulas, 357  
 Frobenius theorem, 133  
 function, 7  
 fundamental system of solutions, 133, 138  
 Fundamental theorem of arithmetic, 571  
  
 game tree, 658  
 games, 658  
 Gauss–Ostrogradsky theorem, 524  
 Gaussian elimination, 78  
 general Fourier series, 427  
 general position, 207, 232  
 generalized binomial theorem, 665  
 generalized exponential power series, 670  
 generates, 94  
 generating function, 664  
 generators, 94  
 geometric basis, 232  
 geometric mean, 288  
 geometric multiplicity, 115, 163  
 geometric points, 230  
 geometric series, 296  
 Gibbs phenomenon, 425  
 gradient of the function  $F$ , 498  
 Gramm determinant, 219  
 Gramm–Schmidt orthogonalisation process, 106  
 graph, 622  
 greatest common divisor, 567  
 Green's theorem, 523  
 Gronwall's inequality, 536  
  
 Hölder's inequality, 448  
 Hölder's inequality for integrals, 450  
 Hamiltonian cycle, 640  
 harmonic mean, 288  
 head, 622  
 Heaviside's function, 258  
 Hermite's interpolation polynomial, 256  
 Hermitian matrices, 160  
 high pass filter, 428  
 Homogeneous coordinates, 229  
 homogeneous coordinates, 230  
 homogeneous equations, 529  
 Homogeneous linear difference equation of order  $k$ , 138  
 homogeneous linear recurrence, 138  
 homogeneous system, 132  
 homothety, 109  
 hyperbolic functions, 300  
 hypercube, 624  
 hyperplane in  $A_n$ , 207  
 hypothesis, 21  
  
 identity relation, 38  
 Image, 99  
 immersions, 514  
 impartial, 660  
 implicit description, 205  
 improper, 267, 277  
 improper integrals of the first kind, 371

- improper integrals of the second kind, 371
- improper limit points, 267
- incident, 622
- inclusion-exclusion principle, 19
- increasing at the point, 279
- increasing on the interval, 279
- indefinite, 228, 487
- indegree  $\deg_+ v$ , 627
- independent variable, 7
- index, 584
- induced subgraph, 626
- infimum, 259
- infinite points, 230
- infinite series of numbers, 291
- infinitely dimensional, 96
- ingoing, 622
- initial values, 525
- injective, 38
- integral curve, 541
- integral domain, 6
- integral mean value theorem, 374
- integral of the form  $\omega$  on  $M$ , 518
- integral operators, 438
- interior, 447
- interior point, 264
- interior point of a subset, 456
- interpolation polynomial, 251
- interval  $[x_0, x_1]$ , 257
- invariant subspace, 112
- inverse, 77
- inverse Fourier transform, 439
- inverse function, 282
- inverse matrix to the rotation matrix, 32
- inverse relation, 39
- inversion in permutation  $\sigma$ , 85
- invertible matrix, 77
- isolated point, 265, 456
- isometry, 453
- isomorphism, 625
  
- Jacobi symbol, 600
- Jacobi theorem, 228
- Jacobian matrix of the mapping, 488
- Jarník's algorithm, 652
- Jordan blocks, 164
- Jordan curve theorem, 645
- Jordan decomposition, 164
- Jordan measure, 374
  
- k-combinations, 15
- k-combinations with repetitions, 15
- kernel of linear mapping, 99
- kernel of the integral operator  $L$ , 438
- Kronecker delta, 75
- Kruskal's algorithm, 651
- Kuratowski theorem, 646
  
- L'Hospital's rule, 285
- Lagrange algorithm, 226
- Lagrange interpolation polynomial, 253
- Lagrange's mean value theorem, 284
- Laplace expansion, 90
- Laplace transform, 443
- law of cosines, 215
- law of inertia, 227
- law of quadratic reciprocity, 597, 598
- leading principal minors, 89
- leading principal submatrices, 89
- leaf, 640
- least common multiple, 567
- left-sided limit, 268
- Legendre polynomials, 418
- Legendre symbol, 595
- Leibniz criterion, 294
- Leibniz rule, 280
- length of a curve, 374
- Leslie model for population growth, 146
- level sets, 498
- limes superior, 294
- limit, 261, 267
- limit point, 456
- limit point of the set  $A$ , 263
- limit points of a subset  $A \subset X$ , 447
- line segment, 208
- Linear algebra, 24
- linear approximation, 255
- linear combinations, 82
- linear combinations of vectors, 24
- linear difference equation of first order, 9
- linear form  $\eta$  on  $U$ , 514
- linear forms, 103
- linear functionals, 435
- linear mapping, 28
- linear mapping (homomorphism), 99
- linear programming problem, 134
- linear restrictions, 134
- linearly dependent, 82
- linearly independent, 82
- Lipchitz continuous, 459
- Lipschitz continuity, 493
- local parametrization of the manifold, 515
- locally finite cover by parametrizations, 518
- logarithmic function with base  $a$ , 276
- logarithmic order of magnitude, 547
- loop, 622
- low pass filter, 428
- lower bound, 259
- lower Riemann sum, 366
- Lucas's test, 609
  
- Möbius function, 579
- Möbius inversion formula, 579
- Malthusian population growth, 9
- mapping, 7
- mapping from a set  $A$  to the set  $B$ , 37
- Markov chain  $x_n$ , 151
- Markov process, 151
- mathematical analysis, 249
- mathematical induction, 10, 14
- matrices, 28
- maximum, 485
- mean value, 374
- member of the determinant 84
- Menger's theorem, 634
- Mersenne primes, 573
- method of Lagrange multipliers, 501
- metric, 445
- metric on the graph, 635
- metric space, 445
- minimum, 485
- minimum excluded value, 661
- minimum spanning tree, 651
- Minkowski inequality, 449

- minor, 89
- minor complement, 89
- modules over rings, 73
- Monte Carlo methods, 24
- morphism, 625
- multidimensional interval, 503
- multiple, 566
- multiplicative function, 580
- mutually perpendicular, 161
  
- natural logarithm, 276
- natural spline, 257
- negative definite, 228
- negative semidefinite, 228
- negatively definite, 487
- negatively semidefinite, 487
- neighborhood of a point, 264
- Newton integral, 359
- nilpotent, 163
- Nim, 658
- non-homogeneous linear difference equations, 143
- Norm, 148
- norm, 155, 445
- norm of the partition, 365
- normal space, 500
- normal vector, 498
- normalised, 104, 153
- normalized vectors, 31
- normed vector space, 445
- nowhere dense, 451
- number  $\pi$ , 297
- number of solutions of a congruence, 588
  
- objective function, 134
- odd, 85
- One-sided derivatives, 277
- one-to-one, 38
- onto, 37
- open, 446
- open  $\varepsilon$ -neighbourhood, 446
- open cover, 265, 456
- open intervals, 264
- open set, 264
- order of  $a$  modulo  $m$ , 582
- order of an integer modulo  $m$ , 582
- order of magnitude, 547
- ordered field, 259
- ordered trees, 643
- ordering, 39
- orientation, 36, 218
- orientation of the manifold, 517
- oriented euclidean (point) space, 218
- oriented manifold with boundary, 522
- oriented manifolds, 518
- oriented vector space, 218
- origin of the affine coordinate system, 204
- orthogonal, 104, 153
- orthogonal basis, 104
- orthogonal complement, 105, 154
- orthogonal group, 157
- orthogonal mapping, 112
- orthogonal matrices, 113, 157
- orthogonal mother wavelet, 427
- orthogonal system of functions, 419
- orthogonally diagonalisable, 161
- orthonormal basis, 153
- orthonormal system of functions, 419
  
- orthonormalised basis, 104
- oscillates, 291
- osculating circle, 353
- outdegree  $\deg_v$ , 627
- outer product, 220
- outgoing, 622
  
- pairwise coprime, 570
- parametric description, 26, 205
- parametrized by the length, 356
- parent, 642
- Parity of permutation, 85
- Parseval equality, 156
- Parseval's theorem, 420
- partial derivatives of order  $k$ , 481
- partial derivatives of the function  $f$ , 476
- partial sums, 291
- particular solution, 133
- partition, 40
- Pascal triangle, 15
- path, 623, 625
- path graph of length  $n$ , 623
- path of length  $n$ , 625
- perfect numbers, 573
- periodic, 298
- periodic function, 422
- permutation, 12
- permutation of the set  $X$ , 84
- permutation with repetitions, 15
- perpendicular, 31, 104, 161
- perpendicular projection, 105
- Perron-Frobenius theory, 147
- Petersen graph, 624
- phase frequency, 425
- Picard's approximation, 532
- planar graph, 645
- plane trees, 643
- Pocklington-Lehmer, 609
- points, 203
- polar basis, 226
- polynomial order of magnitude, 547
- polynomials, 8
- positive definite, 227
- positive direction, 32
- positive matrix, 147
- positive semidefinite, 228
- positively definite, 163, 487
- positively semidefinite, 163, 487
- power function  $x^a$ , 275
- power mean with exponent  $r$ , 288
- power residue, 594
- power series, 295
- power series centered at  $x_0$ , 300
- predecessor, 642
- preimage, 38
- primality witness, 609
- prime, 570
- Primitive matrix, 147
- primitive root, 584
- principal matrices, 89
- principal minors, 89
- principle of inclusion-exclusion, 20
- private key, 612, 614
- probability function, 17
- projection, 105
- projective maps, 231

- Projective plane  $\mathcal{P}_2$ , 229
- projective quadric, 235
- projective transformations, 231
- projectivization of a vector space, 230
- proper, 267, 277
- proper rational functions, 363
- pseudoinverse matrix, 175
- pseudoprime, 605
- public key, 612, 614
- pullback of the form  $\eta$  by  $\varphi$ , 515
  
- QR decomposition, 175
- quadratic forms, 223
- quadrics, 223
  
- Rabin cryptosystem, 613
- radius of convergence, 295
- range, 7
- rank of the matrix, 82
- rank of the quadratic form, 224
- ratio of points, 212
- rational functions, 274
- rays, 208
- real-valued functions of a real variable, 249
- recurrence relation, 9
- reduced residue system, 581
- reflection through a line, 33
- reflexive, 39
- regular collineations, 231
- regular square matrix, 77
- residual capacity, 655
- Riccati equation, 529
- Riemann integral, 365
- Riemann measurable, 504
- Riemann measure of the set, 373
- Riemann sum, 365
- Riemann–Stieltjes integral sum, 386
- right-continuous or left-continuous, 272
- right-sided limit, 268
- Rolle’s theorem, 284
- root of the polynomial, 251
- root of the tree, 641
- root vector, 165
- rooted trees, 641
- rotation or curl of the vector field, 524
- rows of the matrix, 74
- RSA, 612
  
- Saarus rule, 84
- sample space, 17
- sampling interval  $\tau$ , 443
- scalar functions, 7
- Scalar product, 104
- scalar product, 31, 74, 153
- scale, 36
- second-order partial derivatives, 481
- self-adjoint, 160
- self-adjoint matrices, 160
- semiaxes, 225
- semipath, 655
- separated variables, 370
- sequence  $a_n$  converges to  $a$ , 268
- sequentially continuous, 369
- series of functions, 295
- set of solutions, 588
- shift of the plane, 25
- signature of a quadratic form, 227
  
- simplex, 208, 209
- Simpson’s rule, 386
- sine Fourier series, 427
- singular values of the matrix, 173
- sink, 654
- size of a flow, 654
- size of the vector, 153
- smooth, 342
- solution, 525
- solvable, 136
- source, 654
- spanning subgraph, 626
- spanning tree, 649
- spectral radius of matrix  $A$ , 148
- Spectrum of linear mapping, 115
- spectrum of linear mapping, 163
- Sprague–Grundy function, 661
- Sprague–Grundy theorem, 662
- square matrix, 75
- square wave function, 425
- Standard affine space  $\mathcal{A}_n$ , 202
- standard basis of  $\mathbb{K}^n$ , 98
- standard maximalisation problem, 134
- standard minimisation problem, 134
- standard unitary space, 153
- stationary point of the function, 485
- stationary points, 344, 501
- Steinitz exchange theorem, 97
- Steinitz’s theorem, 648
- stochastic matrices, 152
- stochastically independent, 20, 21
- strategy, 658
- strict extremum, 485
- subdeterminant, 89
- subgraphs, 626
- submatrix of the matrix  $A$ , 89
- subspaces, 208
- successor, 642
- sum of impartial games, 661
- sum of subspaces, 95
- supremum, 259
- surface and volume of a solid of revolution, 376
- surjective, 37
- symmetric, 39, 86
- symmetric bilinear form, 108
- symmetric mappings, 160
- symmetric matrices, 160
- symmetrization, 622
  
- tail of the edge, 622
- tangent hyperplane, 480
- tangent line, 255
- tangent line to the curve  $c$ , 476
- tangent plane, 480
- tangent space, 500
- tangent space  $TU$ , 514
- tangent vector, 476, 513
- Taylor expansion with a remainder, 345
- Taylor polynomial of  $k$ -th degree, 345
- the backward difference, 358
- the central difference, 358
- the class of functions  $C^{k_c}(A)$ , 342
- the curvature of the curve, 356
- the differential of function  $f$ , 352
- the differentiation of the second order, 358
- The domain of the relation, 37

- The Euclidean plane, 30
- the existence of a neutral element, 5
- the existence of a unit element, 5
- the existence of an inverse element, 5
- the forward difference, 358
- the Frenet frame, 356
- The fundamental theorem of algebra, 343
- the graph of a function, 38
- the indefinite integral, 358
- the integral mean value theorem, 369
- the lower Riemann integral, 369
- the main normal, 356
- the primitive function, 358
- the second derivative, 342
- the sources and sinks of the network, 653
- the uniform continuity, 369
- the upper Riemann integral, 369
- the Weierstrass test, 382
- topology, 264, 447
- topology of the complex plane, 264
- topology of the metric spaces, 447
- topology of the real line, 264
- torsion of the curve, 356
- totally bounded, 457
- Trace of mapping, 111
- trace of matrix, 111
- trail, 625
- transformation, 489
- transient, 152
- transitive, 39
- translation, 25, 203
- transpose, 86
- transposition, 84
- trapezoidal rule, 385
- tree, 640
- triangle, 208, 623
- triangle inequality, 155
- trigonometric functions, 297
  
- unbounded, 264
- undirected graph, 622
- uniform continuity, 368
- uniformly bounded, 460
- uniformly Cauchy, 380
- unit decomposition subordinate to a locally finite cover, 519
- unit matrix, 32, 75
- unitary group, 157
- Unitary isomorphism, 154
- unitary mapping, 154
- unitary matrices, 157
- Unitary space, 153
- universal formula, 667
- unsaturated, 655
- upper bound, 259
- upper Riemann sum, 366
  
- Vandermonde determinant, 252
- variation, 387
- vector, 72
- vector field, 541
- vector field  $X$ , 513
- vector field  $X$  along the curve  $M$ , 513
- vector functions, 354
- vector functions of one real variable, 354
- vector of restrictions, 134
- Vector space, 92
- vector subspace, 94
  
- vectors, 24
- vertices, 622
  
- walk, 625
- walk of length  $n$ , 625
- wavelet mother function, 427
- weak connectedness, 633
- weakly connected, 639
- weight, 636
  
- zero curvature, 352
- zero matrix, 74
- zero measure, 388
- zero vector, 24

**Based on the earlier textbook:**  
**Matematika drsně a svižně**  
**Jan Slovák, Martin Panák, Michal Bulant**  
**a kolektiv**

published by Masarykova univerzita in 2013  
1. edition, 2013  
500 copies  
Typography,  $\LaTeX$  and more, Tomáš Janoušek  
Print: Tiskárna Knopp, Černčice 24, 549 01 Nové Město nad  
Metují